Modeling Mosquito Activity Built on Mosquito Population Dynamics:

A Simulation Study

A Thesis Submitted to the College of
Graduate and Postdoctoral Studies
In Partial Fulfillment of the Requirements
For the Master of Science degree
In the Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon

By

Peibo Cong

Permission to Use

In presenting this thesis in partial fulfillment of the requirements of the master of statistics degree from the department of Mathematics and Statistics of the University of Saskatchewan, I agree that the Libraries of the University of Saskatchewan may make it freely available for inspection. I also agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other uses of materials in this thesis in whole or part should be addressed to:

Head of the Department of Mathematics and Statistics

Room 142 Mc Lean Hall

University of Saskatchewan

Saskatoon, SK S7N 5E6

Canada

OR

Dean

College of Graduate and Postdoctoral Studies

University of Saskatchewan

116 Thorvaldson Building, 110 Science Place

Saskatoon, Saskatchewan S7N 5C9 Canada

Abstract

Background: West Nile virus (WNv) continues to be one of the most destructive mosquito borne diseases in the world, and Saskatchewan has experienced the highest incidence rates for WNv in North America. Its primary transmitters are mosquitoes, with *Culex tarsalis* serving as the main vector in Saskatchewan. For this reason, mosquito population dynamics is an important determinant of WNv risk. Weather factors, in turn, exert a pronounced impact on mosquito populations. It is important to understand the environmental factors playing a crucial role in oscillations of the mosquito population. It is also important to construct a model or create a method which can monitor and accurately estimate the overall dynamics of the mosquito population.

Methods: In this study, a Probability Generating model is developed to simulate the mosquito observation counts, making use of a pre-existing System Dynamics Model to simulate a mosquito population. A MCMC method was further used to draw samples from a posterior distribution for Bayesian inference and analyse how frequency of observation of mosquito trap counts can improve performance of our model or method.

Purpose of study: This study mainly focuses on investigating the feasibility of estimating the regression coefficients of the logistic regression model for the parameters (β) by using the proposed computational method. Meanwhile, we consider comparing the performance of this method with analysis under different sampling frequencies.

Results: The results of the Probability Generating model depicts the distribution of the simulated observation data (y_i) over our study region (city of Saskatoon) seasonally, which suggests the environmental variables have a significant effect in driving variations in mosquito populations under the simulation experiments; the results of the three different sampling frequencies suggest that the current frequency (weekly) of measuring counts of trapped mosquitos is insufficient for reliable estimation of the parameters (β) for the durations examined.

Conclusion: In this study, we formulated a probabilistic model from a combination of a reasonably complex dynamic model and a probabilistic generating model. Additionally, we have

investigated the frequency of collecting real-world data associated with the accuracy of the model and revealed the importance of sampling mosquito population every day for reliably estimating parameter values, rather than pursuing the standard of sampling mosquito population every week.

Keywords: West Nile virus (WNv); System Dynamics Model (SDM); Probability Generating model; Markov Chain Monte Carlo (MCMC); Environmental Variables; Highest Posterior Density (HPD)Interval.

Acknowledgements

After finishing the thesis, I would like to appreciate the people who have supported and helped me so much throughout this entire process. Without their help and guidance, in order to learn what I needed to, I would have had to take numerous detours. Writing this article has had a significant impact on me, not only on my knowledge, but also challenged me as well.

I would first like to thank my supervisors, Professor Juxin Liu and Professor Nathaniel Osgood. The doors of Professors Liu and Osgood office were always open whenever I had a question on my research or writing and they always gave me a valuable feedback. They consistently allowed this thesis to be my own work, but guided me in the right the direction whenever they thought I needed it.

I would like to thank Winchell Qian, who was involved in editing the code of experiments for this research project. He gave me lots of help on how to edit the code of experiments and on mining data. Without his passionate participation and input, the excellent results of experiments could not have been successfully conducted. I need to thank the environment Canada for providing me the mosquito and environmental data.

I would also like to thank for the professors and staff at the Dept. of Mathematics and Statistics for helping me finish my thesis.

Finally, I must express my very profound gratitude to my family for providing me with unfailing support and encouragement throughout my years of study and through the process of researching and writing this thesis. Thank you.

Table of Contents

Permission to Use……..……..……..……..……....**i**

Abstract.....**ii**

Acknowledgements……...……**iv**

Table of Contents……..……..……..……**v**

List of Tables……..……...**ix**

List of Figures…….……**x**

Chapter 1 Introduction……..……...……..……1

1.1 Why Estimate the Mosquito Population?.....1

1.1.1 Background..……..……...……..……1

- **Definition**
- **History**
- **Risk in Saskatchewan**

1.1.2 Importance of Estimating the Mosquito Population…….…..……..3

- **Mosquitoes as vectors**
- **The size of the vector population is important in mosquito control**

1.1.3 Value of Relatively Precise and Timely Estimation of Mosquito Population…...……..…….……4

1.2 Existing Methods for Estimating the Mosquito Population Size…….……..…..5

- **Mark-Release-Recapture (MRR) experiment**
- **The Fisher-Ford Method**
- **Logistic Regression Model based on MRR**
- **Bayesian Hierarchical Model based on MRR**

- **Dynamic Hydrology Model**
- **System Dynamics Model**

1.3 Overview of Proposed Method…….……..7

1.3.1 Brief Data Description…….…….…..……..7

- **Weather data**
- **Simulation data**

1.3.2 Construction of Study Methods…….…....8

1.3.3 Advantages of SDM and Markov Chain Monte Carlo (MCMC)…….....9

Chapter 2 System Dynamics Models and Markov Chain Monte Carlo …….....10

2.1 Anylogic……...…….……..10

2.1.1 Definition of Anylogic……..…….…….…….10

2.1.2 Anylogic Terminology……..…….…….10

2.2 Generative Model……..…….……12

2.2.1 Introduction of SDM……..…….……13

2.2.2 Generating Simulation Data through Generative Model….……..14

2.3 Methodology: Markov Chain Monte Carlo (MCMC)……..…….15

2.3.1 Introduction of MCMC….……....15

- **What is Markov Chain Monte Carlo?**
- **What is a Markov Chain?**
- **What is Monte Carlo?**

2.3.2 Algorithms of MCMC……17

- **Metropolis-Hasting Algorithm**
- **Gibbs sampling**

2.3.3 Why Bayesian MCMC? 20

Chapter 3 Proposed Method……..…….…….21

3.1 Assumptions …….……...21

3.2 Definitions and Formulas ……..……...…….21

3.3 Concerns in the Model……..……..…….22

- **Beta values**
- **Link function**
- **Standardized data**

Chapter 4 Experimental Design and Results…….…….…..……24

4.1 Experimental Variables and Parameters…….……24

4.1.1 Dependent Variables……..…….……24

4.1.2 Independent Variables…….…….…24

4.1.3 Parameters…….……..…25

4.2 Experimental Statistic Distribution Framework……26

4.2.1 Binomial Distribution……..…….……..26

4.2.2 Logistic Regression Model……..…….……..26

4.2.3 Normal Distribution…….……..…….27

4.3 Experimental Strategy and Procedure…….……29

4.3.1 Experimental Strategy…….…….…….…29

4.3.2 Experimental Procedure…….....…….31

4.4 Experimental Results……..…....…..34

4.4.1 Experimental Results of the Generative Model…….....34

4.4.2 Experimental Results of MCMC…….…….....35

4.4.3 Experimental Results from a Sensitivity Analysis……...41

Chapter 5 Discussion and Conclusion…….……..…50

5.1 Discussion…….……..50

5.2 Potential Future Work…….…….....52

5.2.1 Important Variables…….……..……52

5.2.2 Interactions of Independent Variables…….…….....54

5.3 Conclusion……..55

References….....57

List of Tables

Table 1.1 West Nile virus cases in Canada……..……1

Table 1.2 Saskatchewan Human WNv neuro invasive cases and deaths 2003 – 2014…….……..2

Table 4.1 Coverage rate of the HPD intervals under three scenarios……47

List of Figures

Figure 1.1 WNv transmission cycle……..…..2

Figure 2.1 Stock and Flows for a simple mosquito population model….…….……11

Figure 2.2 Generative Model.....…….....14

Figure 4.1 Experimental statistical framework…….…….28

Figure 4.2 Experimental strategy……..…….…30

Figure 4.3 Stock and Flow Structure of the Mosquito Population Model…….....32

Figure 4.4 The output of running the Generative model……..…….……34

Figure 4.5 Density plots…….……35

Figure 4.6 Trace plots……36

Figure 4.7 Autocorrelation plots……37

Figure 4.8 Running Mean…….….38

Figure 4.9 Diagnostic tests……..…40

Figure 4.10 Mosquito counts in three scenarios……43

Figure 4.11 Scenario 1 sampling period based on 1 day…….…..45

Figure 4.12 Scenario 2 sampling period based on 3 days……..…….…..46

Figure 4.13 Scenario 3 sampling period based on 7 days…….…46

Figure 5.1 The change of number of generated mosquitoes (y_i) from 2010 to 2013…….……..50

Chapter 1 Introduction

1.1 Why Estimate the Mosquito Population?

1.1.1 Background

Definition: West Nile virus (WNv) is an arbovirus belonging to the genus Flavivirus in the family Flaviviridae. Symptoms of WNv infection include skin rash, fever with muscle ache, and sometimes encephalitis or meningitis [1]**.** WNv is spread especially from birds to humans by mosquitoes.

History: In 1937, the first patient with West Nile virus was identified in the West Nile district of Northern Uganda in Africa [2]. Since that time, WNv has spread rapidly throughout Africa and regions of the Middle East, the prevalence of WNv in children was as high as 8%. WNv has become an endemic disease in these countries and regions [3]. Since the mid-1990s, numerous epidemics have also occurred in Europe. In 1999, the first North American case of WNv was reported in Queens, New York City in the United States. After this time, many serious cases of encephalitis were found surrounding the borough of Queens. In the fall of 1999, WNv was declared an endemic disease in the United States [4].

Canada had its first confirmed infection in a bird in 2001. In September 2002, the first confirmed human cases of WNv were reported in parts of Quebec and Ontario. In 2003, WNv cases were reported in four-fifths of the provinces (British Columbia, Alberta, Saskatchewan, Manitoba, Ontario, Quebec, New Brunswick, Nova Scotia) and one-third of the territories (Yukon Territory) in Canada [5]. Since 2003, there have been WNv cases each year in Canada. From Table 1.1 below, we can see that WNv is still active in Canada.

Table 1.1: West Nile virus cases in Canada

Year	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
No. of case	414	1481	25	225	151	2215	36	13	5	101	428	115	21	78

Source [6]: from http://healthycanadians.gc.ca

In 2003, the first case was found with WNv infection in a person from Saskatchewan. Table 1.2 below demonstrates the major outbreaks of WNv occurred in 2003 and in 2007 in Saskatchewan, resulting in serious adverse health outcomes.

Table 1.2: Saskatchewan Human WNv neuro invasive cases and deaths 2003 – 2014

Year	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Neuro invasive Cases	63	0	6	3	76	1	0	0	0	0	7	1
Deaths	7	0	3	0	6	0	0	0	0	0	1	0

Source [7] : from West Nile Virus (WNv) Surveillance Results and Transmission Risk 2015 by the Ministry of Health.

Risk in Saskatchewan: Saskatchewan had an endemic of WNv in 2003, constituting the most cases in Canada in that year. The provincial health department had received reports of 1080 cases of WNv, in which human cases accounted for 947 cases, accompanying 133 equine cases [8]. Seven patients died because of the infection (see Table 1.2). In 2006, the Five Hills Health Region of Saskatchewan reported human seroprevalence of WNv at 9.98%, the highest record in North America [9]. In 2007, a serious outbreak of WNv occurred in Canada. In this outbreak, the cases reported by Saskatchewan were 58.01% of Canada (1285 out of 2215) [10], six Saskatchewan persons were killed by WNv. This suggests that there is a large potential risk to Saskatchewan people. There have been 157 severe neurological cases and 17 deaths in Saskatchewan from 2003 to 2014 (see Table 1.2).

Figure 1.1 below on the WNv transmission cycle [11] suggests that mosquitoes are the primary transmitters of WNv, transmitting the virus between birds and humans. The beginning point of the cycle starts in birds with WNv which can serve as a reservoir for WNv. Because infected birds reach high viral loads, when mosquitoes, such as

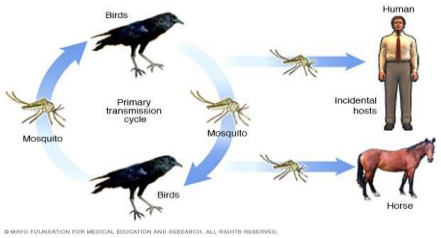

Figure 1.1 (Taken from [11]): WNv transmission cycle

Culex tarsalis, bite the infected birds and suck blood, the mosquitoes can become a vector for WNv. When the infected mosquito bites uninfected birds, the virus enters the bloodstream of those birds. This leads to infection of some birds. Therefore, there is a bidirectional transmission between the birds and the mosquitoes. By contrast, there is one direction between the infected mosquitoes and humans or horses, and the humans and horses are "dead end hosts" representing the end point of the cycle.

In terms of spreading WNv, mosquitoes act as a "bridge" between birds and humans. Therefore, mosquitoes serve as a key factor for controlling WNv. However, not all species of mosquitoes are likely to carry WNv.

1.1.2 Importance of Estimating the Mosquito Population

Mosquitoes as vectors: The mosquito is a member of the family Culicid. Thousands of species of mosquitoes consume the blood of various kinds of animal hosts, including humans. Many species of mosquito can transmit diseases from host to host, where diseases are spread by the bite of an infected mosquito. A mosquito can become infected when it sucks blood from infected animals such as birds and can then spread the pathogen to humans and other animals when it bites. Some mosquito-borne diseases and infections are extremely harmful to humans, such as malaria, yellow fever, dengue fever, West Nile virus, Zika virus, etc.

Mosquitoes as disease vectors have the ability to infect and lead to the death of more humans than any other organism on the Earth – thousands of people die from mosquito borne diseases each year. Mosquitoes carry diseases that affect humans and also transmit many diseases and parasites that affect other animals such as dogs and horses [12]. Mosquitoes spread the diseases which have killed more persons than all the wars in human history [13]. Despite advances in medicine, even just malaria infects tens of thousands of people each year [13]. Therefore, many species of mosquitoes act as "virus super vectors" of diseases. Based on the background and purpose of this study, we are focusing on *Culex tarsalis,* as this species is the principle vector for WNv in Saskatchewan. This mosquito species prefers to breed in newly created freshwater sources, i.e., ditches, standing water pools, etc. During the daytime, adult

Culex tarsalis prefer to seek out shaded areas; however, they are most active in the early morning and dusk and tend to bite birds and mammals.

The size of the vector population is important in mosquito control. The size of the vector population is one of the primary indicators and risk factors with respect to epidemics of mosquito borne diseases. The distribution and amount of the principal vector are central factors in controlling the epidemic of vector-borne pathogens [14]. The size of the mosquito population determines the endemic range and epidemic severity. Measured abundance of the vector provides an indicator of the relative number of mosquitoes in an area during a particular sampling period and can be useful for comparing to thresholds in vector management and in monitoring the outcome of mosquito control efforts. However, greater abundance of mosquitoes does not mean a higher prevalence of WNv. In practice, an epidemic of WNv will follow if these three conditions are met: the amount of mosquitoes is low, percentage of adult mosquitoes over total population is higher and the prevalence of infection among such mosquitoes is high [15]. For example, in 2003, summers in Saskatchewan were very hot and very dry, therefore, many residents thought mosquitoes were very rare. However, approximately 90 percent of mosquito populations were *Culex tarsalis* (the main vector of WNv) and there existed a high proportion of mosquitoes which were infected with the virus (infection rate), so there was an outbreak of WNv. Seven patients were killed by WNv in that year [16].

1.1.3 Value of Relatively Precise and Timely Estimation of Mosquito Population

In Saskatchewan, historical events involving WNv and physiological characteristics of *Culex tarsalis* have shown that it is the primary vector of WNv in Saskatchewan. In order to effectively control WNv in Saskatchewan, WNv monitoring is reported on a weekly basis through the summer. A relatively precise and timely estimation of the *Culex tarsalis*' population is valuable. The reasons are as follows.

(1) The distribution and amount of mosquitoes exerts a strong influence on the epidemic spread of WNv [14];

(2) In order to achieve effective control of WNv vectors in a specific district, predicting the amount of mosquitoes across the district based on sampling from a restricted area is necessary [14];

(3) Precise population prediction of *Culex tarsalis* could provide sufficient time for predicting WNv occurrences, to initiate disease control and start public health interventions. Therefore, the prediction of mosquito population is important in assessing the risks of WNv.

(4) Predicting and tracking the abundance of *Culex tarsalis* is a primary task for surveillance and control programs in a health region [17];

(5) Accurate evaluation of the size of the vector population is a critical factor for understanding the ecology of the vector, and also to plan effective vector control activities [18].

Better estimation of the size of mosquito population can provide valuable information for decision making for government and health authorities. For example, making a reasonable financial budget for WNv disease and creating an optimal medical inventory for disease control could provide enough warning time to prevent an epidemic occurrence and timely control of the further development of WNv disease outbreaks.

1.2 Existing Methods for Estimating the Mosquito Population Size

As mentioned previously in 1.1.3, estimating population size is an important task for any epidemic areas of vector-borne pathogens. The literature offers several methods to address it, which include the following:

Mark-Release-Recapture (MRR) experiment [18]: MRR is a method frequently used to estimate the population size (N) of a certain species in ecology. The stages of the experiment are (1) a portion of the natural population is captured (M) and marked in some method; (2) the captured population is then released into the natural population; (3) Another portion is captured (n) and the number of marked individuals within the second portion is counted (m).

Finally, a solution is given based on a simple equation (m/n=M/N) under some assumptions which lie beyond the scope of our research. This solution provides an estimate of population size as follows:

$$\widehat{N} = n * \frac{M}{m} \quad \text{.....(Eq 1.1)}$$

where, N, M, n and m are as described above [18].

The Fisher-Ford Method [19]: Fisher and Ford (1947) gave detailed information regarding this method, which has been used in MRR experiments and determines population size using ratios of marked to unmarked individuals. This method has been further developed by Dowdeswell (1959) and Parr (1965). The method has been replaced by modern methods and is now hardly used.

Logistic Regression Model based on MRR. Cianci et al. (2013) applied statistical tools, such as a logistic regression model, with Mark-Release-Recapture experiments to estimate size of mosquito population and assessed the performance and accuracy of this model by using simulated data from known population sizes [18].

Bayesian Hierarchical Model based on MRR. Villela et al. (2015) constructed a hierarchical probabilistic model and performed a Bayesian analysis using this model to estimate the mosquito population using data from MRR experiments. Using the Bayesian analysis by Markov Chain Monte Carlo method, an inference concerning the size of mosquito population was obtained. To get a precise estimation of mosquito population abundance, the researchers implemented multiple runs of MCMC via the Gibbs sampling algorithm, and then the results were given with properties of statistical measures. In the process of applying the model, authors used the JAGS model and WinBUGS tool in order to get the estimates [20].

The Bayesian Hierarchical Model has a high accuracy for estimating mosquito populations compared with the Fisher-Ford method [20].

There are some other models or methods for estimating population size that don't use the data from the MRR experiment:

Dynamic Hydrology Model. Shaman et al. (2002) used a dynamic hydrology model with time series regression analysis, making it possible to predict different species of mosquitoes' population sizes in a temporal and spatial status. This model was driven by environment variables, including the air temperature, precipitation, relative humidity, wind speed, surface pressure, etc. These environmental factors were treated as influencing the probability of reporting a mosquito in the model. The authors provided a basic path toward a dynamic mosquito prediction system [21].

System Dynamics Model. During the past decade, many researchers have built system dynamics models (SDM) based on life cycles of mosquitoes which were applied in field of diffusion and control of disease transmitted by the mosquito. Brailsford et al. (2008) applied SDM to assess risk of mosquito borne diseases. In their assessment procedure of risks, the mosquito population was estimated indirectly [22]. There were also course projects that used SDM to estimate mosquito populations [23].

For much of the literature review, the principal methods or tools that have been applied in estimating mosquito population include logistic regression, time series analysis, and the Markov chain Monte Carlo (MCMC) method.

1.3 Overview of Proposed Method

1.3.1 Brief Data Description

Weather data: Our time period for weather data extended from 2010 to 2013 in the city of Saskatoon. The data sets were downloaded from the National Climate Data and Information Archive and Environment Canada. From the weather data sets, environment variables such as temperature, humidity, wind speed and precipitation were selected and used to generate three different scenarios by processing the standardized data. The scenarios are one day means, three days means and seven day means, respectively. The procedure created a time series. To understand the time series effects of the above environment variables on subsequent mosquito population size, a lag of means of environment variables could be created.

Although there is missing data for some weather variables, they are below the 5% threshold. On the whole, the weather data satisfies our frequency, availability and accuracy requirements.

Simulation data: Based on the problems with collecting data sets about empirical mosquito counts, we sought to employ a simulation experiment, in which "pseudo empirical" (synthetic empirical) data was generated and used for analysis. In order to generate the pseudo observation data set (y_i), there are three stages: 1) simulate the total (synthetic) mosquito population (N_i) using the System Dynamics Model; 2) calculate the (synthetic) probability of capturing a mosquito (p_i) by using a logistic regression model on the empirical weather data; 3) generate the pseudo number of captured mosquitoes (y_i) by using the Binomial distribution model using probability (p_i).

There are two main advantages to this method: first, it does not affect the purpose of this study, which is to investigate the feasibility of estimating the regression coefficients for the parameters (β) using different frequency of data; secondly, use of the pseudo data sets can avoid the problems with model misspecification and with the empirical data.

1.3.2 Construction of Study Methods

Constructing our model consisted of two stages: looking for a statistical theoretical foundation and proposed applied models. Here, the reference to the term foundation of statistical theory refers to statistical knowledge and principles. For instance, this includes various statistical definitions, theorems and distributions of different kinds. Finding the appropriate foundation of statistical theory was based on the procedure of our experiments. And proposed applied models are built on this foundation of statistical theory under the background of our experiment.

In our experiment, the procedure of captured mosquitoes (a dichotomous outcome) suggested use of a binomial distribution and logistic regression. Here, we assumed that the number of captured mosquitoes follows a binomial distribution, and that the logit of the probability of capturing of a given mosquito can be characterized as a linear function of the environmental variables. We will give a further discussion in later chapters. Based on the first

stage and the background of our experiment, the MCMC method is attractive to apply in this study.

1.3.3 Advantages of SDM and Markov Chain Monte Carlo (MCMC)

Based on the literature review, there are several advantages of using a dynamic model in a simulation experiment: 1) offering continuous real-time estimation of mosquito populations which are currently impossible to measure in practice; 2) expanding the application range of its methods by adjusting the model for different health regions by accounting for environment conditions; 3) providing a powerful tool to public health departments with flexible methods for the estimation of current and future vector population size.

There are also several advantages of using the MCMC method: 1) Probabilistic estimation sample values of parameter estimates from a posterior distribution instead of estimating a single point estimate. The results can specify statistical measures that include means, medians, standard deviations and credibility intervals; 2) Applying the MCMC method in a temporal-spatial modeling can produce more precise estimation [20]; 3) In general, the MCMC method makes our computations easier than the traditional MLE method in handling estimation parameters of our interest when the procedure of estimation involves a complex integral or cannot be integrated.

Chapter 2 System Dynamics Models and Markov Chain Monte Carlo

In order to understand this study, this chapter focuses on the following three aspects: introducing the models which were involved in this study (Anylogic and generative models); describing the MCMC simulation method which is applied in this study; and lastly, proposing the Bayesian logistic regression model and its assumptions. Some issues in the model are illustrated.

2.1 Anylogic

2.1.1 Definition of Anylogic

Anylogic is software for simulation. It is a powerful and useful tool which supports modeling languages and development tool for Discrete Event Modeling, System Dynamics Modeling and Agent Based Modeling simulation methodologies [24].

2.1.2 Anylogic Terminology

The Anylogic simulation language mainly consists of the following constructs [25]:

1). **Stock & Flow Diagrams**: used for System Dynamics model;

Here, we depict the relationship between a stock and its flows using simple mosquito population model, and reveal mathematical meaning behind the relationships. In relation to a given stock, there are two types of flow: inflow and outflow. An example stock and flow diagram for a simple mosquito population model is given in the below:

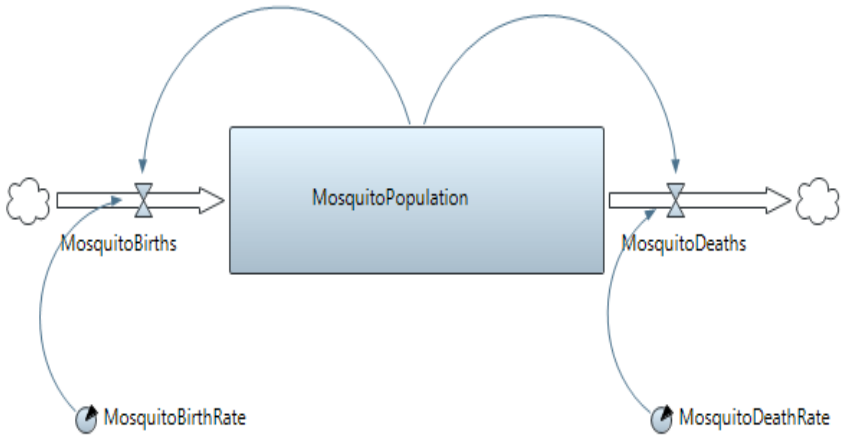

Figure 2.1 Stock and Flows for a simple mosquito population model

In the above figure, the MosquitoPopulation (*N*) is a stock, which is an accumulation of mosquitos associated with time. The mosquito births and mosquito deaths represent an inflow and outflow to that stock, respectively. From the figure, the values associated with the inflow and outflow are the rates at which given mosquito quantity is increased to or decreased from the stock associated with time, respectively. And the MosquitoBirthRate (*BR*) and MosquitoDeathRate (*DR*) are the parameters of this model. These quantities are linked via the following equations (Eq 2.1) and (Eq 2.2).

$$Mosquito\ births = BR \times N \quad \text{.....(Eq 2.1)}$$

$$Mosquito\ deaths = DR \times N \quad \text{.....(Eq 2.2)}$$

Based on the mathematical definition of the derivative, the relationship between the stock variable (MosquitoPopulation (N)) and the flow variables (mosquito births and mosquito deaths) is as approximated in the following equations.

$$N_{t+\Delta t} = N_t + (BR - DR) \times N_t \times \Delta t \quad (0 \leq \Delta t \leq 1) \quad \text{.....(Eq 2.3)}$$

Where the Δt denotes the time step, which can assume a positive value between zero and one.

Equation (Eq 2.3) can be rewritten as following,

$$\frac{N_{t+\Delta t} - N_t}{\Delta t} = (BR - DR) \times N_t \quad \text{.....(Eq 2.4)}$$

When letting $\Delta t \to 0$, we obtain the following equation:

$$\frac{dN_t}{dt} = \lim_{\Delta t \to 0} \left(\frac{N_{t+\Delta t} - N_t}{\Delta t}\right) = (BR - DR) \times N_t \quad \text{.....(Eq 2.5)}$$

The equation (Eq 2.5) is a first-order ordinary differential equation corresponding to the system of the simple mosquito population model.

In general, we can write the relationship between stock and flow as follows:

$$Stock(t) = \int_0^t (inflow(x) - outflow(x))dx + S_0 \quad \text{.....(Eq 2.6)}$$

Where S_0 is the stock at the initial time when $t = 0$.

2). **State charts:** mainly used in Agent Based model to define agent behaviors, and also sometimes used in Discrete Event models as well;

3). **Action charts**: are used to define algorithms, primarily in Discrete Event models and in Agent Based models;

4). **Process flowcharts**: which are the basic construction used to define processes (e.g., defined workflows) in Discrete Event models.

The language also includes other useful objects, including, but not limited to, low level model constructs (variables, equations, parameters, events etc.), presentation shapes (lines, polylines, ovals, etc.), analysis facilities (datasets, histograms, plots), connectivity tools, standard images, and experiments (scenarios).

2.2 Generative Model

In this section, we will introduce the basic ideas behind generating the simulation data that drives the mosquito population (N_i) and observed mosquito number (y_i) through a probability generating model. Within probability and statistics, a generative model is a model which provides a method to randomly generate a set of observed samples.

2.2.1 Introduction of SDM

A System Dynamics model is defined as an approach to build a dynamic model to characterize, simulate, explain, analyze dynamically and manage complex problems or systems in real world applications [26]. At the end of the1950s, Jay W. Forrester invented this approach and in the 1960s this approach obtained further improvements and applications in many fields [26].

The procedure of creating a System Dynamics model mainly includes the following four stages: conceptualization, formulation, testing and implementation. In building System Dynamics Model, modellers use stocks (accumulations) and flows (rates of change), time delays and feedbacks to understand the behaviour of complex systems over time [27].

Why should we choose SDM for generating mosquito counts in this study? The choice of approach should be based on the following two factors: 1) the system being modeled; 2) the purpose of the modeling.

An important background element for this study is the ecosystem of mosquitoes which includes environmental factors, food chains, vegetation, etc. These factors can directly or indirectly affect the temporal and spatial distribution of the mosquito population. The effects of all factors on the abundance of mosquitoes as well as interactions among them form a dynamic system. The mosquito life cycle also is a continually moving process of reproduction, maturation, diapause and death. The purpose of this study is to investigate the problem of estimating of the relationship between environment variables and probability of measuring a mosquito, as would be useful to estimate the size of the mosquito population. In the real world, the size of mosquito population is difficult to accurately measure, which would pose problems for evaluating the accuracy of the estimation procedure and accompanying scenarios examined in this study. But the System Dynamics Model can generate synthetic mosquito populations of known size, which can then serve as reference “ground truth” to assess the accuracy of estimation across different scenarios. It is believed that the relevant mosquito population dynamics characterized by the System Dynamics Model approximately matches those from the real world system in which a mosquito survives, and is thus valuable for this study.

2.2.2 Generating Simulation Data through Generative Model

From the figure below, the mosquito population (N_i) cannot be readily measured in the real world. Moreover, a continuous and complete observation set of data, which are the actual numbers of captured mosquitoes in each period, cannot be obtained from respective health regions because there is an absence of availability and consistency regarding the data.

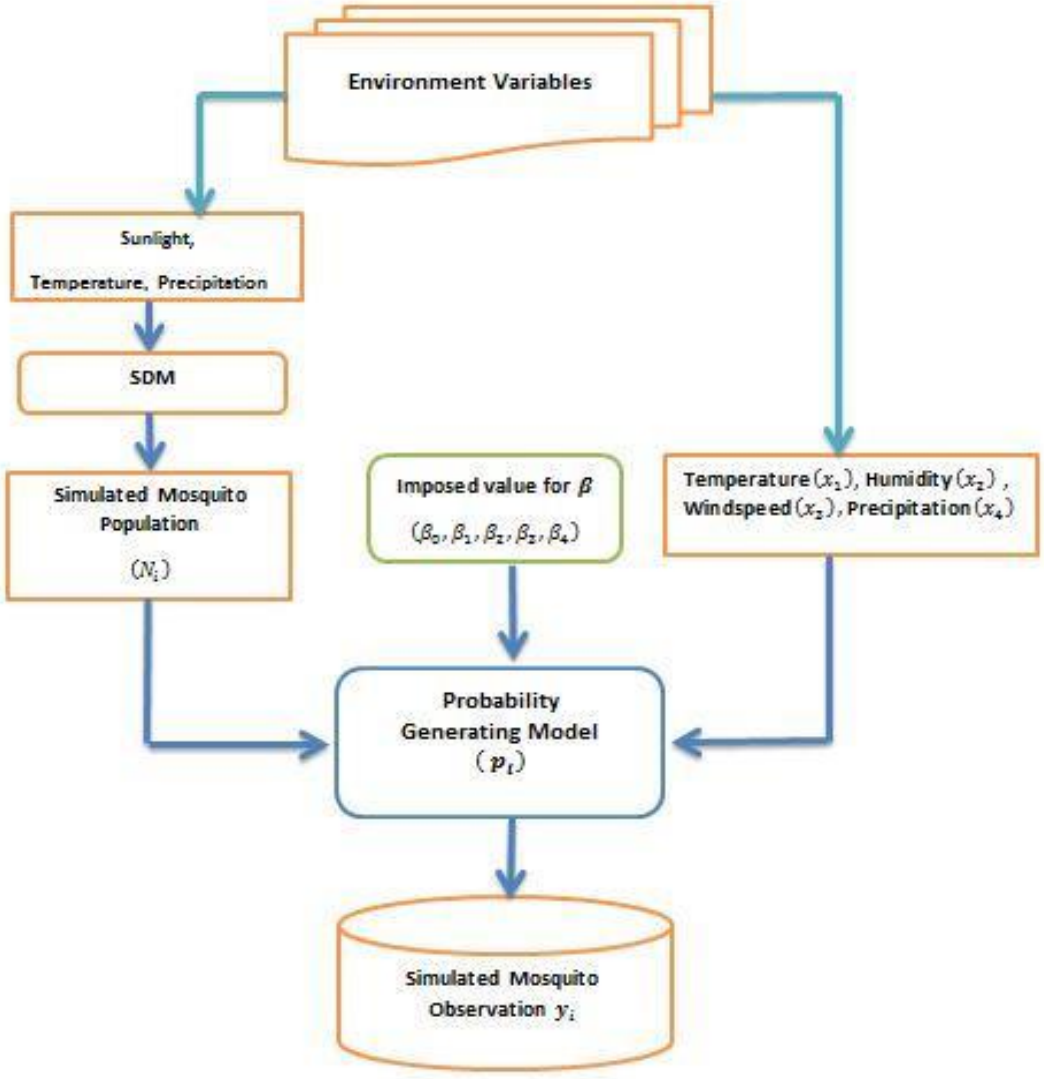

Figure 2.2: Generative Model

In order to realize the goal of this project, we need to simulate these data by constructing a generative model. The use of simulation studies can let us avoid the complications stemming from uncertainties regarding the real-word relationships between the observed number of captured mosquitoes and the population of mosquitoes. In detail, inputting environment data, temperature, precipitation and sunlight, etc., into a System Dynamics Model, the outputs from that model are the pseudo (synthetic) underlying mosquito population (N_i) in each time period. In the approach used here, we then input the environment data and the pseudo mosquito population (N_i) into the probabilistic generative model which was built to generate probabilities of capturing mosquitoes by application of logistic regression. Finally, we obtain the synthetic observed (captured) mosquito count (y_i) in each time period. Our purpose is to test how realistically accurate we can estimate parameters (β) with these synthetic mosquito count under different sampling frequencies (e.g., daily, every three days and every seven days). Here, by representing a known "ground truth" concerning the underlying mosquito population the generative model allows for model evaluation under different sampling regimes.

2.3 Methodology: Markov Chain Monte Carlo (**MCMC)**

2.3.1 Introduction of MCMC

What is Markov Chain Monte Carlo (MCMC)? The MCMC method is a set of algorithms for drawing samples for a probability distribution based on creating a Markov chain with a desired (target) distribution as its equilibrium distribution [28]. MCMC methods are often applied to solve integration and optimisation problems in high-dimensional spaces.

The foundational theorems and tools behind MCMC are Markov Chains, Monte Carlo integration and algorithms which are tools for constructing a Markov Chain. We describe these basic ideas as follows:

What is a Markov Chain? Suppose that we have a stochastic process $\{S_t\}$, where t denotes time. The state space associated with S_t can be a finite set or an infinite set. When $S_t = n_t$, it indicates the process at time t to be in the n_t^{th} state. A Markov Chain can be defined as below:

If a state sequence is $n_0, \ldots, n_{t-1}, n_t, n_{t+1}$, and $t \geq 0$, the stochastic process $\{S_t\}$ satisfies the following requirement,

$$P\{S_{t+1} = n_{t+1} | S_0 = n_0, \ldots S_{t-1} = n_{t-1}, S_t = n_t \} = P\{S_{t+1} = n_{t+1} | S_t = n_t\} \ldots \text{(Eq 2.7)}$$

Then the stochastic process $\{S_t\}$ is called a Markov chain. In fact, a Markov Chain is a special kind of stochastic process [29].

What is Monte Carlo? The Monte Carlo method is a random sampling numerical method based on probabilistic statistical theory. Its basic ideas are: first, linking the problem which will be solved with a certain probability distribution; second, creating a mathematical or statistical model based on the problem; third, repeating a random sampling trial based on the model by using computer simulation technology; and fourth, applying the Law of Large numbers to approximately estimate the solutions of the problem. The procedure of the Monte Carlo method can be seen as a method to calculate integrals by using random sampling trials. For example, suppose a problem for figuring out a complex integration can link to solving an expected value of a random variable $p(x)$ with a probability distribution density function $q(x)$. This can be written down as below.

$$E[p(x)] = \int_0^\infty p(x)q(x)dx \approx \frac{1}{K} \sum_{i=1}^{K} p(x_i) \ldots\ldots\ldots\ldots\ldots \text{(Eq 2.8)}$$

Then, this problem can be solved with following steps:

1) Creating a model which is based on the problem;

2) Repeating a random sampling trial based on running the model, obtaining K simulation values $x_1, x_2, \ldots, x_K$ which are drawn randomly K samples from $q(x)$.

3) Calculating the average value of K random variables' value $p(x_1), p(x_2), \ldots, p(x_K)$. Applying the Law of Large Numbers, we obtain the approximate value of the integral as follows,

$$E[p(x)] = \int_0^\infty p(x)q(x)dx \approx \frac{1}{K} \sum_{i=1}^{K} p(x_i) \ldots\ldots\ldots\ldots\ldots \text{(Eq 2.9)}$$

When the number of sample $K \to \infty$, the estimated value is the actual value of the integral.

2.3.2 Algorithms of MCMC

There are two MCMC algorithms that are easy to implement and broadly applicable. One is the Metropolis-Hasting Algorithm, another one is Gibbs Sampling.

Metropolis-Hasting Algorithm: The Metropolis-Hasting algorithm is one of the most popular MCMC methods and is widely applied in various fields today [30]. It is mainly applied to simulate the samples from troublesome distributions. The algorithm implicitly implements a Markov chain. It follows the stages below:

Suppose our goal is to sample from a target distribution $p(x)$ which is difficult to achieve, but we are given $q(x)$ as a proposal distribution, from which it should be easy to sample.

1). Choose a starting value $x^{(0)}$ which is drawn at random from proposal distribution $q(x)$, with $p(x^{(0)}) > 0$.

2). At iteration n, draw a candidate x^* from the proposal distribution $q\left(x^* \middle| x^{(n-1)}\right)$;

3). Compute the Metropolis-Hasting ratio R $(x^{(n-1)}, x^*)$, where

$$\mathrm{R}\left(x^{(n-1)}, x^*\right) = \frac{p(x^*)q\left(x^{(n-1)}|x^*\right)}{p(x^{(n-1)})q\left(x^*|x^{(n-1)}\right)} \quad \text{.....(Eq 2.10)}$$

4). Sample a value for $x^{(n)}$according to the following steps:

(a) Independently draw a randomly value u from Uniform distribution$(0,1)$;

(b) If $u \leq min\{ R\ (x^{(n-1)}, x^*),1\}$, then accept x^* and set $x^{(n)} = x^*$, otherwise, set $x^{(n)} = x^{(n-1)}$ (i.e., repeat the previous sample).

5). Increase $n\ by\ 1$ and repeat stages 2-4 m times to get m samples from $p(x)$, with optional burn-in periods and discarding samples to achieve a desired thinning ratio.

The terms burn-in period (burn-in) and thinning in step 5 refer to an output of MCMC. A sample description is as follows.

Burn-in: At the beginning of an MCMC simulation, often can see the performance in initial iterations is poor because of the following two reasons: 1) irregular fluctuations due to the initial iterations, which were influenced strongly by starting value $x^{(0)}$; 2) they don't provide much more useful information on the target distribution. Based on the above reasons, this initial part of the iterations often offers little value for any inference. So we discard them as a "burn-in period".

Thinning: Thinning is a technique applied if significant autocorrelation obtains between the observed output samples such as could be determined via the use of autocorrelation plots. Significant autocorrelation is not what we want, as it lowers the effective sample size, which can be deleterious for certain tasks. In order to decrease the autocorrelation associated with samples, a useful method is to thin the Markov chain by holding every i^{th} $(i \geq 1)$ sample from each sequence and discarding the others. This procedure is called thinning.

Gibbs sampling: In statistics, Gibbs sampling is a special case of the Metropolis–Hastings algorithm that is another popular MCMC algorithm for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution, when direct sampling is difficult. We can consider the problem as follows: assume a distribution of interest (target distribution) is $\mathrm{p(x)}$, where x is a vector $= (\mathrm{x_1}, \mathrm{x_2}, \dots, \mathrm{x_k})^{\mathrm{T}}$, and denote $\mathrm{x_{-i}} = (\mathrm{x_1}, \dots, \mathrm{x_{i-1}}, \mathrm{x_{i+1}} \dots, \mathrm{x_k})^{\mathrm{T}}$. Also consider a case where the full conditional distributions $\mathrm{p_i(x_i)} = \mathrm{p(x_i|x_{-i})}$ are available, and easily sample for $i = 1,2, \dots, k$. In general, the procedure of Gibbs Sampling follows these stages:

1) Select initial value $x^{(0)}$, and set $n = 0$.

2) Obtain updated value $x^{(n+1)} = (x_1^{(n+1)}, \dots \dots, x_k^{(n+1)})^T$ from x^n through successive generation of values via sampling from the following distributions:

$$x_1^{(n+1)} \sim p(x_1 | x_2^n, \dots, x_k^n)$$

$$x_2^{(n+1)} \sim p\left(x_2 \middle| x_1^{(n+1)}, x_3^{(n)}, \dots, x_k^{(n)}\right)$$

… … ……..(Eq 2.11)

$$x_{k-1}^{(n+1)} \sim p\left(x_{k-1}\middle|x_1^{(n+1)}, x_2^{(n+1)}, \dots, x_{k-2}^{(n+1)}, x_k^{n}\right)$$

$$x_{k}^{(n+1)} \sim p\left(x_{k}\middle|x_1^{(n+1)}, x_2^{(n+1)}, \dots, x_{k-1}^{(n+1)}\right)$$

3) Increment $n\ by\ 1$ and repeat stage 2).

From the above Gibbs sampling, we can see that Gibbs updates the variable X from the conditional distribution $p\left(x_i\middle|x_1^{(n+1)}, x_2^{(n+1)}, \dots, x_{i-1}^{(n+1)}, x_{i+1}^{(n)} \dots, x_k^{(n)}\right)$. Each Gibbs cycle, which is the completion of step 2 for all components of X, consists of k Metropolis-Hasting steps [31]. To recognize this, one should realize that the i^{th} Gibbs step in a cycle effectively puts forward the candidate vector $X^* = (x_1^{(n+1)}, x_2^{(n+1)}, \dots, x_i^*, x_{i+1}^{(n)}, \dots, x_k^{(n)})$ for the current state of the Markov chain $(x_1^{(n+1)}, x_2^{(n+1)}, \dots, x_i^{(n)}, x_{i+1}^{(n)}, \dots, x_k^{(n)})$[31]. Therefore, the i^{th} Gibbs updates can be seen as a Metropolis-Hasting step sampling

$$X^*|(x_1^{(n+1)}, x_2^{(n+1)}, \dots, x_i^{(n)}, x_{i+1}^{(n)}, \dots, x_k^{(n)}) \sim q_i(x^*|x_1^{(n+1)}, x_2^{(n+1)}, \dots, x_i^{(n)}, x_{i+1}^{(n)}, \dots, x_k^{(n)}) \quad \text{...(Eq 2.12)}$$

Where, $q_i\left(x^*\middle|x_1^{(n+1)}, x_2^{(n+1)}, \dots, x_i^{(n)}, x_{i+1}^{(n)}, \dots, x_k^{(n)}\right) = \begin{cases} p\left(x_i^*\middle|x_{-i}^{(n)}\right) & if\ \ X_{-i}^* = x_{-i}^{(n)} \\ 0 & otherwise \end{cases}$

The Metropolis-Hasting ratio under the above condition is

$$R\left(x_i^{(n)}, x_i^*\right) = \frac{p(x_i^*, x_{-i}^{(n)}) q_i(x_i^{(n)}|x_i^*)}{p(x_i^{(n)}, x_{-i}^{(n)}) q_i(x_i^*|x_i^{(n)})}$$

$$= \frac{p\left(x_i^*, x_{-i}^{(n)}\right) p\left(x_i^{(n)}\middle|x_{-i}^{(n)}\right)}{p\left(x_i^{(n)}, x_{-i}^{(n)}\right) p\left(x_i^*\middle|x_{-i}^{(n)}\right)}$$

$$= \frac{p\left(x_i^*|x_{-i}^{(n)}\right) * p\left(x_{-i}^{(n)}\right) * p\left(x_i^{(n)}\middle|x_{-i}^{(n)}\right)}{p\left(x_i^{(n)}|x_{-i}^{(n)}\right) * p(x_{-i}^{(n)}) * p\left(x_i^*\middle|x_{-i}^{(n)}\right)}$$

$= 1$ ……...(Eq 2.13)

In the above derivation, we applied the following two facts: the proposal distributions $q_i\left(x_i^{(n)}\middle|x_i^*\right)$ for Gibbs sampling are the posterior conditions $p\left(x_i^{(n)}\middle|x_{-i}^{(n)}\right)$ and Bayesian chain rule, the full join distribution equals the product of two terms$\left(e.g., p\left(x_i^*, x_{-i}^{(n)}\right) = p\left(x_i^*|x_{-i}^{(n)}\right) * p\left(x_{-i}^{(n)}\right)\right)$. This indicates that the candidate x_i^* always accepted. Therefore, the Metropolis-Hasting algorithm does the exact same thing as a Gibbs update.

2.3.3 Why Bayesian MCMC?

Bayesian MCMC methods are powerful and useful computational tool for drawing samples from a posterior distribution in Bayesian analysis. The posterior distribution can be expressed as

Posterior distribution $\propto$ *likelihood* $\times$*prior distribution*.....(Eq 2.14)

From this expression, we can note that when parameters (β) of our interest are treated as a random variables. The Bayesian MCMC methods provide a set of samples from the posterior distribution which allow us to derive different inferential statistics (e.g. point estimation, percentile estimation, interval estimation) in a transparent way. Prior distribution allows us to incorporate additional information beyond the observed data (either historical information or information from similar studies). On the other hand, in frequentist statistics, the Maximum Likelihood Estimate (MLE) is a method of arriving at a point estimate for our parameter in interest to find a 'best' value for the parameter so that it maximizes the likelihood function.

In short, the Bayesian MCMC sampling methods provide the samples from the posterior distribution which certainly gives more information than just a single point and interval estimation. This is the main motivation for the use of the Bayesian MCMC method for this work.

Chapter 3 Proposed Method

3.1 Assumptions

Our study area covers the city of Saskatoon. We worked with data collected by capturing adult mosquitoes at seven sites located in this region.

1) We assume that whether or not a given mosquito within the simulated zone is captured is independent (conditional on the value of the covariates) of whether or not another mosquito in that zone will be captured. This assumption is important in the distributional assumptions of our model.

2) Building on the above, we further assume that the process of capturing a single mosquito follows a Bernoulli distribution, and the total number of mosquitoes caught is well-characterized as a Binomial distribution. This assumption indicates that the counts of captured mosquitoes follow a Binomial distribution.

3) The final assumption is that the effects of weather variables on the probability of capturing mosquitoes are independent of each other. This assumption positions that there are not statistical interactions amongst the weather variables in as much as they determine the probability of capturing a given mosquito.

3.2 Definitions and Formulas:

1) Binomial Distribution $Bin(N, p)$: We assume that the process of capturing mosquitoes is a binomial experiment involving the total size of the mosquito population(N) and the probability of capturing each mosquito (p).

2) The probability of capturing a mosquito (p): We assume that the probability of capturing each mosquito is well-characterized by a logistic regression model denoted as

$$\text{logit}\,(p) = ln\,(\tfrac{p}{1-p}) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \beta_4 * x_4 \quad \text{...(Eq 3.1)}$$

Where x_i (i=1,2,3,4) denote the weather variables for temperature, humidity, windspeed and precipitation respectively; β_0 is a intercept term and $\beta_i(i = 1,2,3,4)$ are coefficient of the logistic regression. We can rewrite them as vectors: $X = (1, x_1, \dots, x_4)$ and $\beta = (\beta_0, \beta_1, \dots, \beta_4)^T$, and express the logistic regression model as follows,

$$\text{logit}\,(p) = ln\,(\frac{p}{1-p}) = X\beta \quad \text{.....(Eq 3.2)}$$

3) Rjags: Rjags is a software package that has been introduced in recent years to connect R functions and the Jags library for Bayesian data Analysis [32].

3.3 Concerns in the model

Beta values: as previous mentioned, β is a parameter's vector, and forms a key point of interest in this study. It consists of the intercept term β_0 and the relationships between the weather variables(x_i) and probability of capturing mosquitoes. Its elements consist of β_0, β_1, β_2, β_3 and β_4. We use a Normal prior for each beta parameter, that is $N(0,10000)$. Estimators of these elements will come from sequence samples of its posterior by using the MCMC methods.

The choices of their initial values for β at the beginning of MCMC in simulation studies are flexible. The initial value assumed is important for performance of the model. A better but reasonable selection of their initial values can increase the convergence speed of the Markov Chain.

Link function: as mentioned in previous assumptions, the process of capturing a mosquito is viewed here as equivalent to a Bernoulli experiment. It can repeat again and again. Bernoulli experiment only has possible two outcomes: one is successfully capturing a mosquito; another is failing to capture a mosquito. We sum the individual dependent variables and the sum follows a Binomial distribution. The probability distribution of a Binomial distribution is $p(Y = y) = \binom{N}{y} p^y (1-p)^{N-y}$. Because there is a non-linear relationship between $p(Y = y)$ and the number of captured mosquitoes (y), we need to look for a link function to generate probability $p(Y = y)$ in our generative model. Logistic regression models are commonly used for binary response variables. The link function is as follows:

$$logit(p) = log\frac{p}{1-p} = \eta(x) \text{ ……..(Eq 3.3)}$$

Therefore, our logistic regression model is,

$$log\frac{p}{1-p} = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \beta_4 * x_4 \text{ ……..(Eq 3.4)}$$

We can then derive the inverse logit link function:

$$p(being\ captured|X, \beta) = \frac{e^{\beta_0+\beta_1*x_1+\beta_2*x_2+\beta_3*x_3+\beta_4*x_4}}{1+e^{\beta_0+\beta_1*x_1+\beta_2*x_2+\beta_3*x_3+\beta_4*x_4}} \text{ ……..(Eq 3.5)}$$

Standardized data: We use the standardized weather data instead of raw weather data in the MCMC model. That is we transform environment variables (x) into the form of standardized data by using the z transformation formula below.

$$z_x = \frac{x-\mu_x}{\sigma_x} \text{ ……....(Eq 3.6)}$$

Where μ_x denotes the sample mean of the environment variable values and σ_x denotes the sample standard deviation of the environment variable values. Using the standardized weather data instead of the raw weather data in MCMC model could help speed up the mixing performance of the MCMC and improve the efficiency of the MCMC sampling. In fact, using the raw weather data is reasonable in statistical theory. But with such data, obtaining a good result from running the MCMC model often requires more iterations than with standardized data.

Chapter 4 Experimental Design and Results

What are the effects of environmental variables on number of captured mosquitoes? Identification of a precise and reliable means of measuring these effects is an important goal of this study. The design is based on the study goal and uses several models to support it.

4.1 Experimental Variables and Parameters

4.1.1 Dependent Variables

In this study, the dependent variables which are simulated by using different models and software tools include the following:

- Size of the estimated mosquito population (N)
- The number of captured mosquitoes (y)

Accurate and reasonable simulated values for mosquito population (N) and number of captured mosquitoes (y) will serve as an important enabler for achieving the goal of this study.

4.1.2 Independent Variables

The independent variables in the experiment involve environment variables (weather variables) which have a direct influence on the abundance of mosquitoes or their probability of being captured. For example, studies have shown that temperature is significantly positively associated with the observed mosquito population, while precipitation is negatively correlated with the observed mosquito population [33]. Our original data about weather variables came from Environment Canada [34]. Considering the availability and accuracy of data and other factors, the following variables were selected:

- **Temperature**: Average daily temperature, measured in degrees Celsius (°C), denoted by x_1.

- **Humidity**: Average daily relative humidity. It is a percentage (%) giving the measured partial pressure of water vapor divided by the equilibrium vapor pressure of water under a given temperature condition, denoted by x_2.
- **Windspeed:** Average daily windspeed. Its unit of measurement is kilometer per hour (km/h), denoted by x_3.
- **Precipitation:** Average daily precipitation in millimetres (mm), denoted by x_4.

Time series of these four independent variables used by the thesis was sourced from Environment Canada, with the raw data containing temperature, humidity and precipitation data on a daily basis, while windspeed data is partially daily and partially hourly. We aggregated hourly windspeed data into daily means to unify time series units for all independent variables.

The daily data of these independent variables were transformed by the standardization technique explained in Chapter 3. Following transformation, the transformed values of the independent variables were brought into the analysis.

4.1.3 Parameters

As mentioned above, we had characterized the parameters in a simple way. The parameters consist of an intercept and coefficients of weather variables, which include daily average temperature, daily average relative humidity, daily average windspeed, and daily average precipitation in a logit function $\eta(x)$. The formulation assumes that there are no interactions among the weather variables and that there is a linear relationship between $\eta(x)$ and weather variables. If we denote the intercept constant (or the bias term) as β_0, along with parameters $\beta_1,\ \beta_2,\ \beta_3\ and\ \beta_4$ (as defined in 4.1.2), respectively, parameter vector β can be written as a vector $= (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^T$. In Chapter 3, we assumed that the prior distribution for each of the β_i $(i = 0,1,2,3,4)$ follows a Normal distribution with mean μ_i and variance σ_i^2. For the sake of simulation experiments, we assume a prior for β_i consisting of a Normal distribution with mean $\mu_i = 0$ and variance $\sigma_i^2 = 100^2$. We also mentioned that these parameters are estimated by drawing a sequence of samples from the posterior distribution as discussed in Chapter 3. The following hypotheses are made with respect to the parameters: 1) The probability of capturing a mosquito will increase when the temperature increases within a certain range of

changes; 2) The probability of capturing a mosquito will increase when the relative humidity increases within a certain range of changes; 3) The probability of capturing a mosquito will decrease when the windspeed increases; 4) The probability of capturing a mosquito will decrease when precipitation (or rainfall) increases within a certain range of changes. These general knowledges provide criteria for selecting the sign of the initial values or imposed values of parameters when running the MCMC sampling.

4.2 Experimental Statistic Distribution Framework

To characterize the process of capturing mosquitoes, the framework made use of the following probabilistic and System Dynamics models.

4.2.1 Binomial Distribution

We start by considering a scenario involving capture of a single mosquito; this situation is similar to a coin toss experiment. We assume that capturing a single mosquito follows a Bernoulli distribution $Ber(p)$, where p represents the probability of successfully capturing a mosquito. Z_j denotes the outcome of capturing a single mosquito in trial j and is a dichotomous random variable, with a value of either success (capturing one mosquito) or failure (capturing zero mosquitoes). The probability mass function $f(z_j, p)$ of the Bernoulli distribution, based on the above outcomes for Z_j, is

$$f(z_j, \mathrm{p}) = \begin{cases} p & if\ z_j = 1 \\ 1-p & if\ z_j = 0 \end{cases} \quad \text{.....(Eq 4.1)}$$

Assume that the process of capturing each mosquito is independent and identical. Then the $Z_1, Z_2, \dots, Z_N$ are independent identical random variables, all following the Bernoulli distribution with probability p. Then we can infer that $Y_i = \sum_{j=1}^{N} Z_j$ follows a Binomial distribution $Bin(N, p)$. Its probability mass function can be denoted as follows,

$$f(y_i, N, p) = \binom{N}{y_i} p^{y_i}(1-p)^{N-y_i} \quad \text{.....(Eq 4.2)}$$

4.2.2 Logistic Regression Model

The above probability mass function is used to produce synthetic (pseudo) data about y_i in the probability generating model. In order to do this, we need the size of mosquito population

(N) and the probability of capturing mosquitoes (p). So the Mosquito Population Model is applied to simulate the size of the mosquito population (N) across a time period. The p is defined as a cumulative distribution function for a logistic distribution of $\eta(x)$, which is derived from the following link function,

Step1 $$logit(p) = log\ (\frac{p}{1-p})\ = \eta(x)$$

Step2 $$\frac{p}{1-p} = exp(\eta(x))$$

Final step $$p = \frac{exp(\eta(x))}{1+exp(\eta(x))}$$

Where $\eta(x)$ is a linear function of weather variables under the five dimensional space as mentioned above. Its form is as below,

$$\eta(x) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \beta_4 * x_4 \text{ ……..... (Eq 4.3)}$$

Where $x_1,\ x_2,\ x_3\ and\ x_4$ denote the weather variables: average temperature, average relative humidity, average windspeed and average precipitation, respectively.

4.2.3 Normal Distribution

In the linear function, we assumed a prior for the intercept (β_0) and coefficients of weather variables ($\beta_i, i = 1,2,3,4$) given by a normal distribution $N(\mu_i = 0, \tau_i = \frac{1}{\sigma_i^2} = 0.0001)$.

Linking all statistical distributions pertaining to this experiment, we designed a statistical framework for this experiment. It is depicted in the following figure:

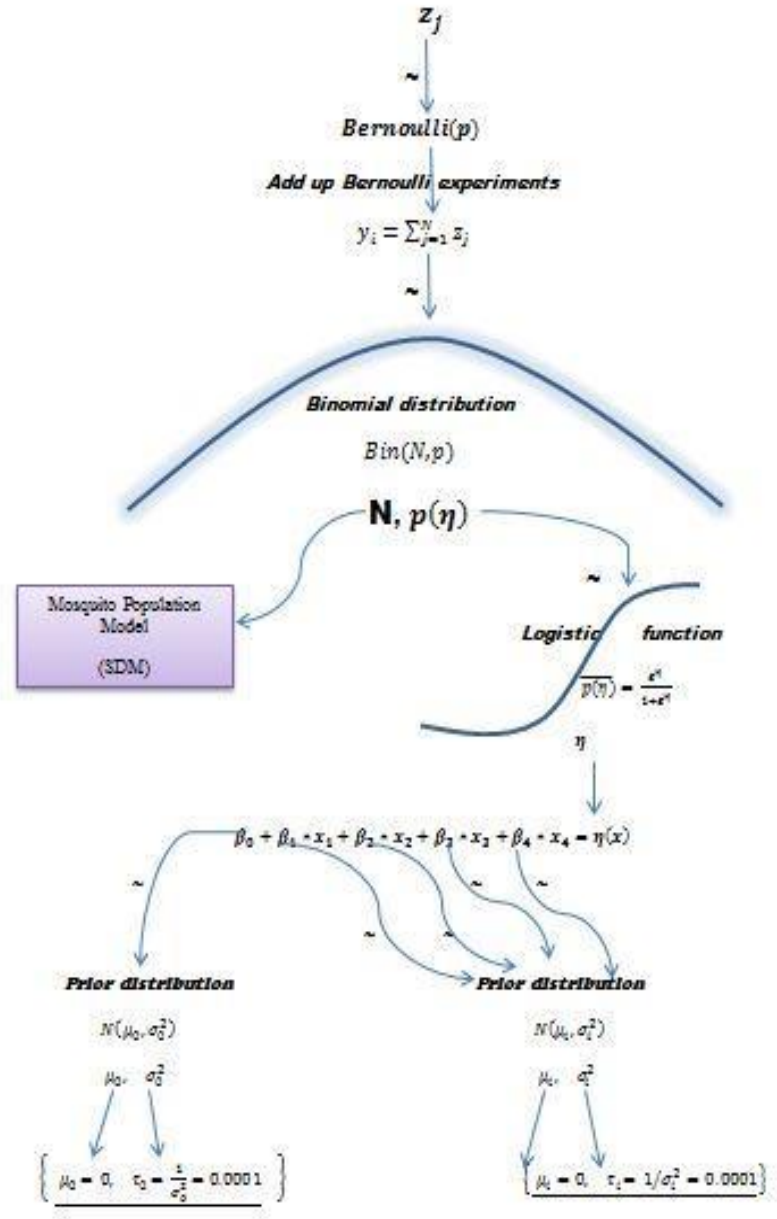

Figure 4.1: Experimental statistical framework

4.3 Experimental Strategy and Procedure

4.3.1 Experimental Strategy

In the last section, we explored the statistical theories which are assumed to characterize the process of capturing mosquitoes, and discussed their role with details in the experiment. In the current section, we will focus on an experimental strategy which must be closer to the goal of the experiment. A reasonable and feasible experimental strategy and logic are a guide or core of the experimental procedure. The below graph is a strategy diagram for this study.

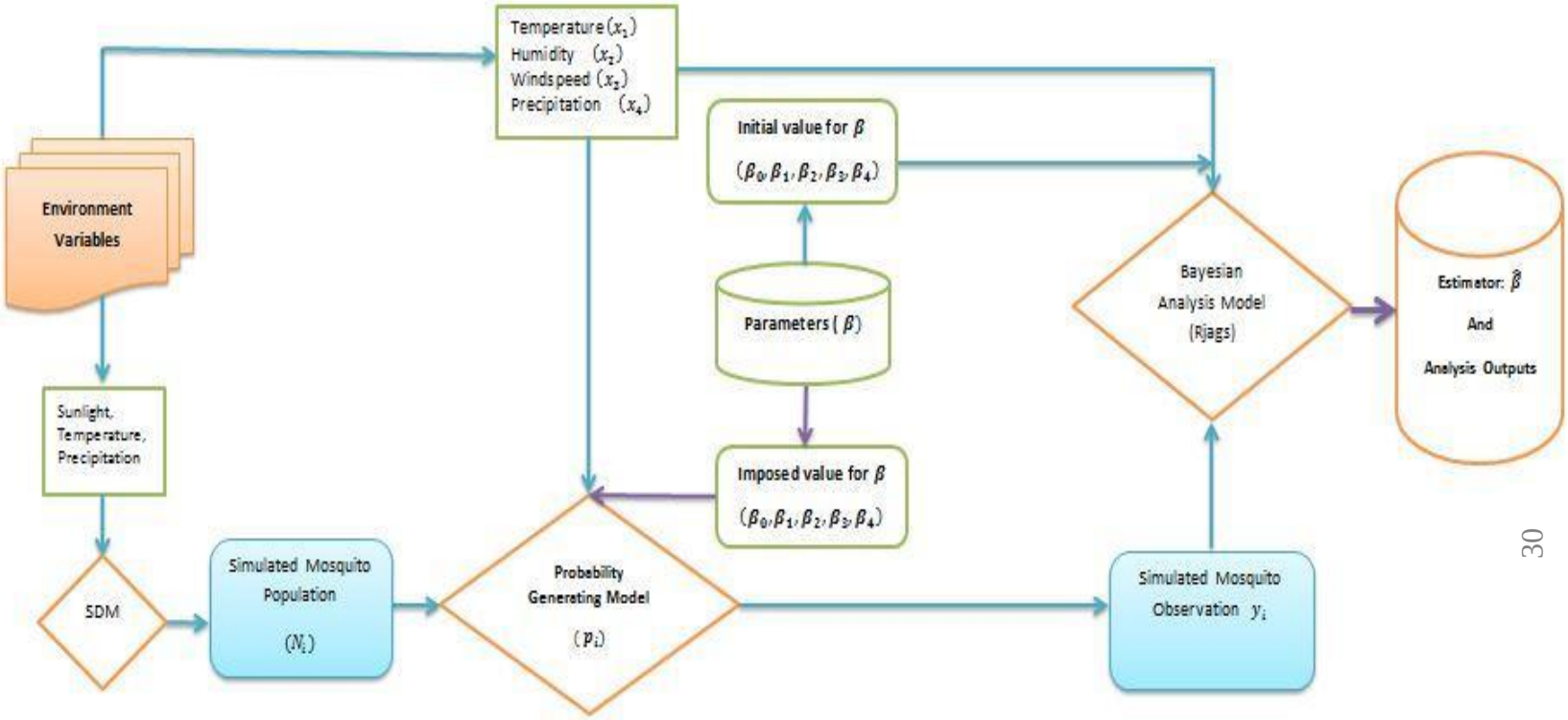

Figure 4.2: Experimental strategy

4.3.2 Experimental Procedure

Based on the above strategy diagram, this experimental procedure follows the below steps, which start with collecting data and finish with parameter estimates.

Step 1: Collect data pertaining to daily weather variables.

Step 2: Handle data, including cleaning data, dealing with missing data, and standardization of data under the following three scenarios: one day, three days and seven days' sampling intervals.

Step 3: Simulate the size of the mosquito population(N_i) on a daily basis: Enter environmental variable data into the Mosquito Population Model (a System Dynamics model in AnyLogic). Data includes daily temperature data, precipitation data, sunlight data, etc. Run the Mosquito Population Model, recording model output. Transform the results into a suitable file structure. It bears emphasis here, that the Mosquito Population Model (a SDM) is quite complex and grounded model characterizing the life cycle of mosquitoes (see Figure 4.3) and its dependence on environmental factors, where the stocks represent state variables, and the flows define rate of change for the stocks in time, in other word, the flows collectively represent the derivatives to those state variables. Therefore, the mathematics underlying the Mosquito Population Model is first-order ordinary differential equations (see Eq 2.5) characterizing how various mosquito subpopulations (as well as the entire such population) change with time. In fact, the SDM is actually characterizing a nonlinear system of first-order ordinary differential equation. The Mosquito Population Model stock–and–flow diagram is as follows:

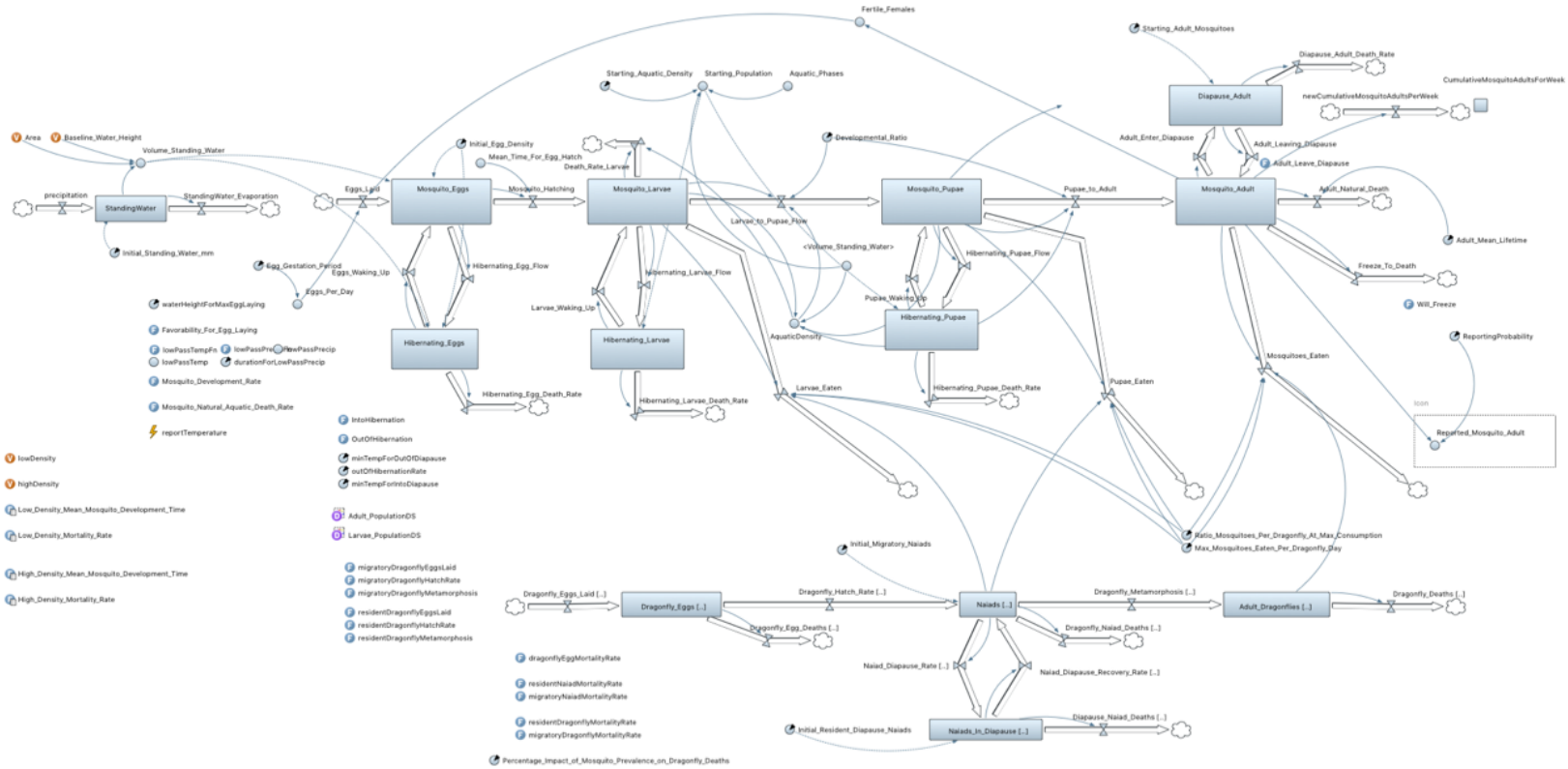

Figure 4.3: Stock and Flow Structure of the Mosquito Population Model [23]

Step 4: Simulate the probability of capturing a mosquito: enter the weather data and N_i into the logistic-regression based probability generating model with the aim of simulating a synthetic (pseudo) probability. The environmental variables consist of temperature, relative humidity, windspeed and precipitation. The mosquito data is the N which came from step 3.

Step 5. Generate synthetic counts of captured mosquitoes based on daily information and then aggregate them for different observation frequencies. This step first generates numbers of daily captured mosquitoes with a random seed and results of step 3 and step 4, and then aggregates the counts of daily captured mosquitoes generated with the same random seed using different frequencies (daily, every three days, or every seven days). These time series of the sum of captured mosquitoes with different aggregation frequencies were used as synthetic (pseudo) observations of captured mosquitoes (y_i), with one particular frequency being used for each experiment.

In this process of step 4 and step 5, the statistical models (logistic regression model and Binomial distribution) seem to be so simple, but there are reasons for this. First of all, the process of capturing a mosquito imply the statistical meaning of logistic regression which is a reasonable and useful statistical tool to calculate the probability of capturing a single mosquito in step 4; Second, one of the reasons that the statistical model can be simpler is that a large subcomponent of the process of generating synthetic counts of captured mosquitoes (y_i) namely, the dynamics associated with the mosquito population (N_i) and its dependence upon several environmental factors was captured within the SDM descried in step 3, thereby, allowing the statistical model to focus simply on characterizing the probability that a given mosquito within the general population will be caught by a trap.

Step 6. Set up the initial value for parameters: $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$.

Step 7. Estimate parameters: import necessary data which come from step 2, step 3 and step 6 into the MCMC model and then draw samples for β from its posterior distribution by running the model under different sampling frequencies.

Step 8. Compare and analyze the posterior samples of β under the three scenarios.

4.4 Experimental Results

4.4.1 Experimental Results of the Generative model

In Chapter 2, we discussed the Generative model. The main purpose of creating the Generative model is to simulate the numbers of captured mosquitoes(y_i). Figure 4.4 depicts the

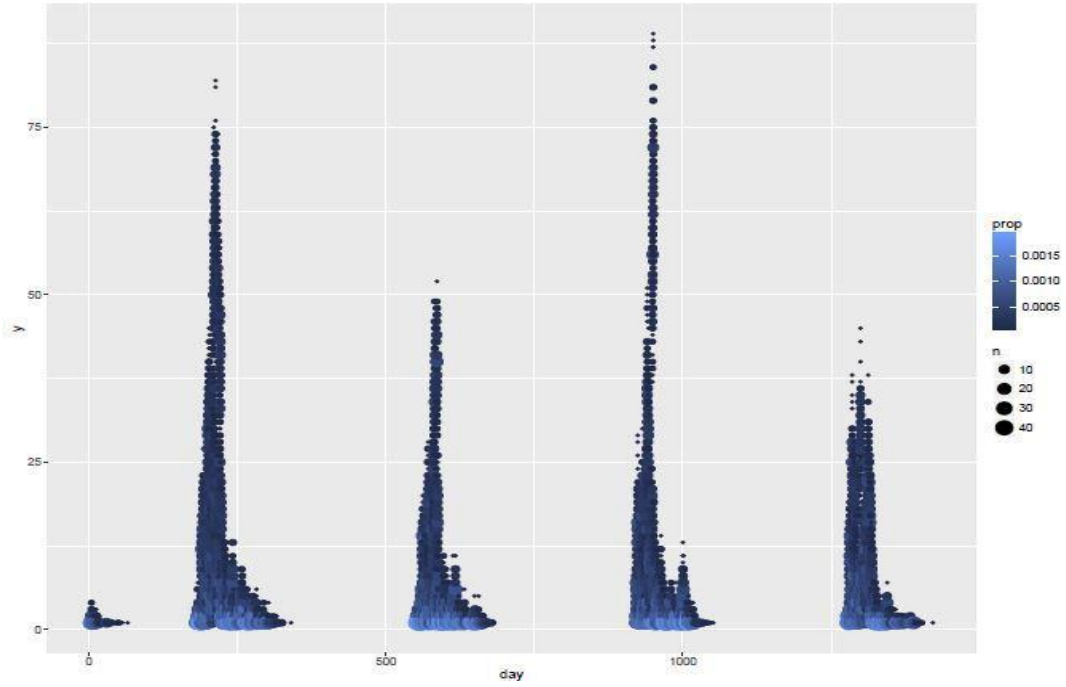

Figure 4.4: The output of running the Generative model

output of running the Generative model on a daily basis. The horizontal axis denotes the daily time from January 1, 2010 to December 31, 2013. The vertical axis indicates the number of captured mosquitoes. There are four major peaks which correspond to the summer season of each year. This suggests that there is a relationship between the simulated numbers of captured mosquitoes and environment variables which are reflective of the actual relationship. This helps build face plausibility for the hope that the Generative model is reliable and useful.

4.4.2 Experimental Results of MCMC

Carrying out the experiment according to the above experimental procedure yields the following outputs for the posterior samples of parameters (β) based on the daily sampling scenario:

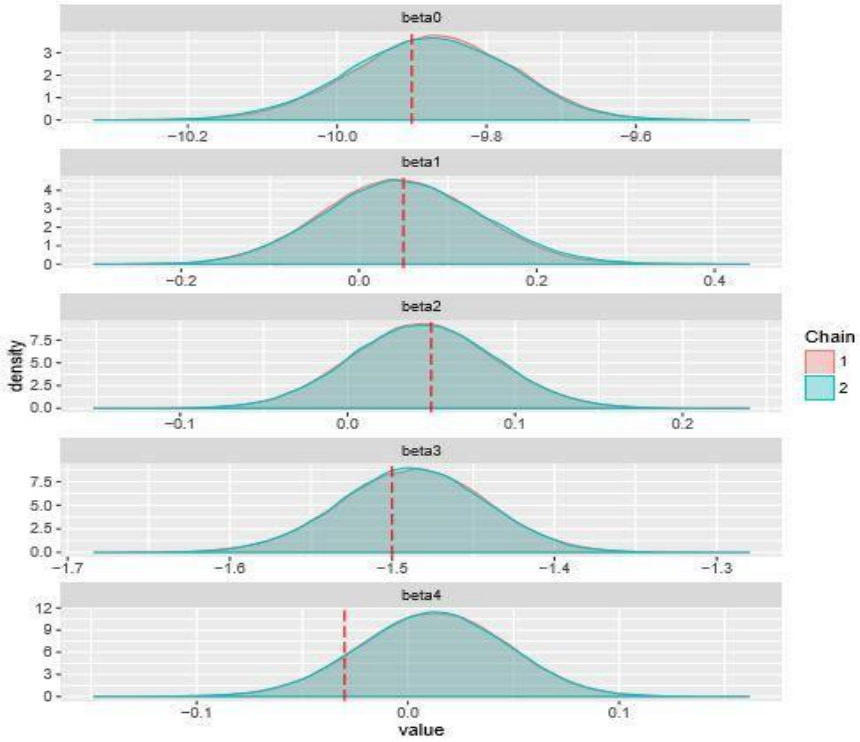

Figure 4.5: Density plots

Explanation: the above curves are posterior density curves, which are estimated from the posterior samples. By showing the posterior density curve, we seek to compare different chains with respect to whether they converge to the same target distribution over time. In this output, there are two overlapped density curves, each associated with a different initial value, which is indicated by a different colour.

Comments: The above density plots for each parameter $\beta_i (i = 0,1,2,3,4)$ indicate that the two different chains converge to the same target distribution. Their shapes are similar to a bell with a single symmetric peak (unimodal), which suggests that target distribution for the

posterior distribution of parameters (β) may be characterized by a normal distribution. One should also note that the estimated mean of the distributions exhibits a slight deviation from the true value of the betas (shown via red dotted lines), especially for β_4 (corresponding to the average of precipitation).

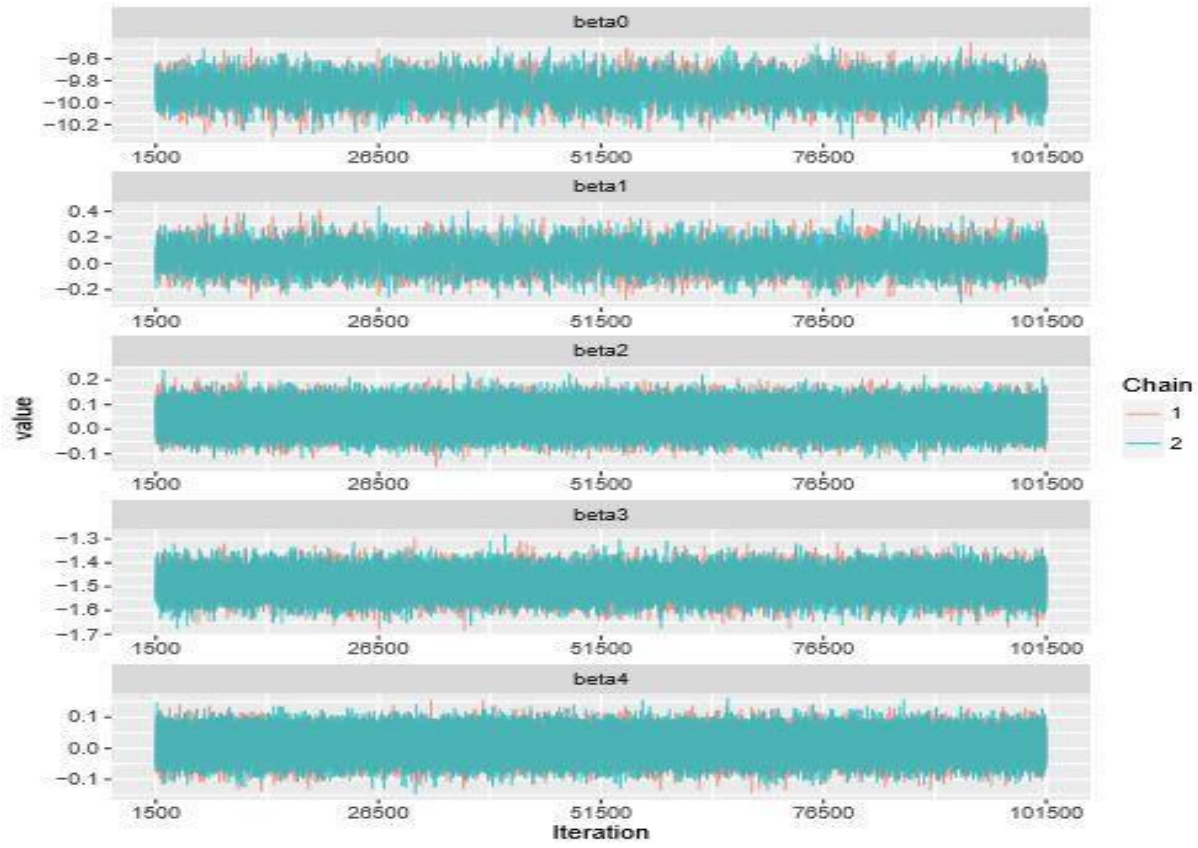

Figure 4.6: Trace plots

Explanation: The trace plot shows the trajectory of the chain, and is applied to check whether the chains converge to same target distribution. In Figure 4.6, after a burn-in period, there are not extreme outlier values for each chain.

Comments: From the above trace plots, we can see that there are two differently coloured lines, which represent two different chains. The two chains didn't get "stuck" at any points or regions in each parameter β_i state space. These plots demonstrate smooth, stable, well-balanced graphs. Although they have some volatility in the initial phase, as the time series extend they show a common trend which converges to the target distribution. However, we need to point out that the two chains didn't mix as well for parameter β_0 and $\beta_1$when compared with other parameters.

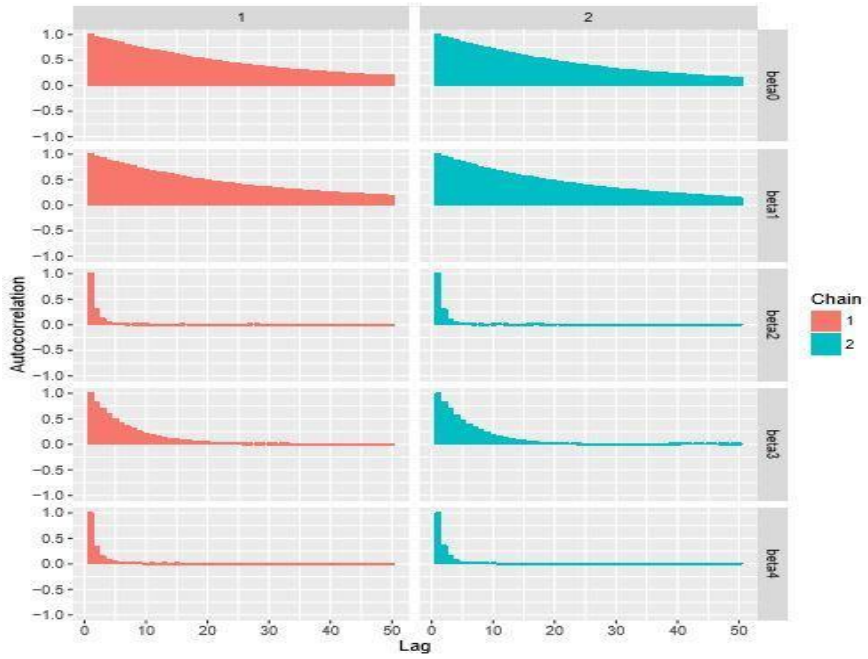

Figure 4.7: Autocorrelation plots

Explanation: Autocorrelation is a measurement which indicates the degree of dependence of the successive samples drawn from the posterior distribution within each chain.

The higher the autocorrelation, the more dependent the samples.

Comments: From the autocorrelation plots, there are higher autocorrelations in the Markov chains about parameters β_0 and β_1 than other parameters; this indicates that the Markov

chains about parameters β_0 and β_1 contain some dependent samples which didn't provide us meaningful information from the posterior distribution. Therefore, these samples reduced the efficiency of these chains and their convergence speed. Meanwhile, we also noted that the autocorrelations exhibited a smooth decrease with some slight vibrations and approached zero as the time and length of chains were extended.

However, Markov chains under other parameters exhibit a relatively small autocorrelation. They quickly get close to zero as the time goes on since the index sample increases. This reveals that these chains are sampling much more efficiently.

Figure 4.7, the autocorrelation plot suggests that the MCMC model is appropriate for this study.

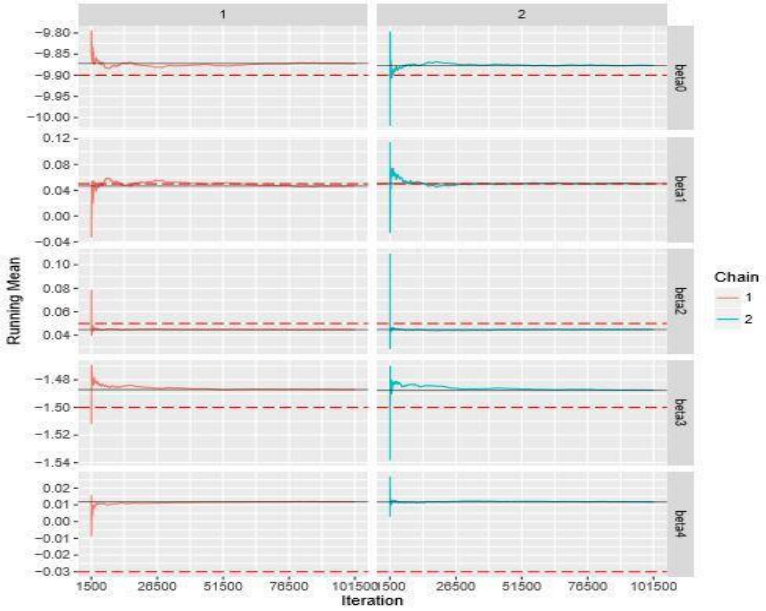

Figure 4.8: Running Mean

Explanation: Figure 4.8 depicts a time series of the running mean; the red dashed lines denote the true values of the parameters (β). Its functions are: 1) checking whether the speed of convergences to a target distribution is slow; 2) comparing if all chains exhibit the same mean when estimating the same parameter.

Comments: The above output clearly shows that the means of two chains are different for parameter β_0. They are all bigger than the true value (-9.9) and they have a vibration around

their mean line on the initial time interval. This indicates that the two chains do not have the exact same mean and that their convergence speeds are both slow.

The mean lines of the two chains for β_1 are on a same mean line and all means are as same as the true value (0.05) with increased iterations; they also exhibit a vibration around their mean line in the initial time interval. This also indicates that the two chains have the exact same mean but their convergence speeds are both slow.

For parameterβ_2, the mean lines of the two chains are lying on the same mean line and are smaller than the true value (0.05). It is also of note that the two chains have no obvious vibration around their mean line on the initial time interval. This suggests that the two chains have the same mean and their convergence speeds are both fast.

The case of the parameter β_3 is an interesting one. The mean lines of the two chains almost lie on the same mean line and higher than the line of the true value (-1.5). There are distinctions between the two chains: chain one has a slight vibration around its mean line in the initial time interval, while chain two has an obvious vibration around its mean line. This indicates that the two chains have approximately the same mean, and that the second chain converges more slowly than the first one.

Concerning parameter β_4, the mean lines of the two chains lie on the same mean line, itself higher than the line of the true value (-0.03). Also we noted that there are very slight vibrations for all two chains around their mean lines in the initial iterations. All of these factors imply that the two chains converge to the same mean line and their convergence speeds are approximately the same.

Overall, the four graphs for each of the chains appear to have consistent traits for each parameter. These consistent traits include converging to the same target distribution for different chains and (with the exception of β_3) convergence speed.

The above comments based on the output graphs reflect certain perceptual judgments. In order to give the performance of the Markov chains a more objective judgement, we analyzed its behaviour using the Heidelberger-Welch diagnostic method (depicted in Figure 4.9).

```
[[1]]

       Stationarity start     p-value
       test         iteration
beta0  passed       1         0.1718
beta1  passed       1         0.0722
beta2  passed       1         0.7652
beta3  passed       1         0.2334
beta4  passed       1         0.0573

       Halfwidth Mean     Halfwidth
       test
beta0  passed    -9.8721 0.005050
beta1  passed     0.0465 0.003935
beta2  passed     0.0447 0.000430
beta3  passed    -1.4871 0.000951
beta4  passed     0.0119 0.000362

[[2]]

       Stationarity start     p-value
       test         iteration
beta0  passed           1     0.749
beta1  passed           1     0.958
beta2  passed           1     0.233
beta3  passed       10001     0.267
beta4  passed           1     0.147

       Halfwidth Mean     Halfwidth
       test
beta0  passed    -9.8769 0.004982
beta1  passed     0.0504 0.003934
beta2  passed     0.0448 0.000415
beta3  passed    -1.4879 0.000959
beta4  passed     0.0118 0.000369
```

Figure 4.9: Diagnostic tests

Explanation: From the output of the Heidelberger and Welch diagnostic, we note that this output includes two tests; a stationarity test and a half-width test. The stationarity test evaluates the stable state of the two Markov chains by testing the null hypothesis that the created two Markov chains have stabilized; the other test assesses if each Markov chain drawing with its associated sample size from the posterior is sufficient to meet the required accuracy for estimating the mean of parameters (β). On the whole, if the Heidelberger and Welch diagnostic test fails, it may imply the Markov chain needs to run for a longer period or highlight some problematic issues with MCMC convergence.

Comments: checking the above output, both Markov chains passed the two tests. Such results suggest that the two Markov chains are stationary and smooth processes. And all chains passed the half-width test. These reflect the fact that drawing the given count of samples from the posterior is enough to reach a pre-specified accuracy for the mean estimate.

In the first Markov chain, we should note that although the chain for parameter β_4, which is the coefficient associated with the precipitation variable, passed the stationarity test

after first iteration, its $p-value$ is only 5.73%. This means that the p-value is close to the significance cut-off (a threshold $\alpha = 5\%$). If the p-value is less than the significance cut-off (for example $\alpha = 5\%$), then we should reject the null hypothesis that the chain is a stationary distribution which is a probability distribution that remains unchanged in the Markov chain as time progress.

4.4.3 Experimental Results from a Sensitivity Analysis

Sensitivity: The above results come from executing our model, which is constituted by many different components. Each component has some effect on the results. But the effects cannot be quantized precisely. Therefore, the outputs/results should be analyzed for sensitivity empirically by observing the results of changing our model elements. In our sensitivity analysis, we changed the sampling frequency of the observed data in MCMC and mainly focus on how the statistical inference results would be affected by different frequencies of the observed data.

Highest Posterior Density (HPD) interval: In contemporary work, a 100(1-α)% HPD interval is a popular method to summarize some important statistical features of posterior distribution for the parameters of interest in Bayesian inference. Observations of samples' posterior distribution with our case study suggest that the distributions are unimodal. We therefore confine our discussion in the following to unimodal posterior density functions. In this context, a simple definition of Highest Posterior Density (HPD) interval is given by following [35]:

A 100(1-α)% HPD interval for β is simply defined by $R(\pi_\alpha) = \{\beta: \pi(\beta|D) \geq \pi_\alpha\}$,

Where, $\pi(\beta|D)$ is the posterior distribution density function of the parameters (β) of interest, given data set D, π_α is the largest constant such that $P(\beta \in R(\pi_\alpha)) \geq 1 - \alpha$.

From the above definition, we can see that an 100(1-α)% HPD must satisfy the following three requirements [35][36]: (1) the posterior probability of the region is 100(1-α)%; (2) inside the interval, the posterior density for every point is greater than every point outside the interval; (3) the interval is the shortest length for a given probability 1-α.

The definition of Highest Posterior Density of the multi-modal distribution is beyond our research. Readers interested in a more detailed definition can consult (Fadallah, A. 2011).

Results of Sensitivity analysis: This sensitivity analysis is based on changing the sampling period for the time series to observe the reaction of the synthetic numbers of captured mosquitoes y_i and the estimated parameters (β) under the three different sampling period scenarios: 1-day, 3-days and 7-days. The y_i are the results output from the Generative model; the parameters (β) are the results sampled from the posterior distribution $\pi(\beta|D)$. These results are analyzed by visual inspection and quantitative analysis, and using the HPD interval diagnostic method.

1. **Results from the Generating model under three different scenarios**

Figure 4.10 depicts results showing mosquito counts over time from performing the Generative model under the three scenarios: scenario 1 = 1-day sampling period, scenario 2 = 3-day sampling period and scenario 3 = 7-day sampling period. For scenarios 2 and 3, the relative mosquito counts y_i is the daily average.

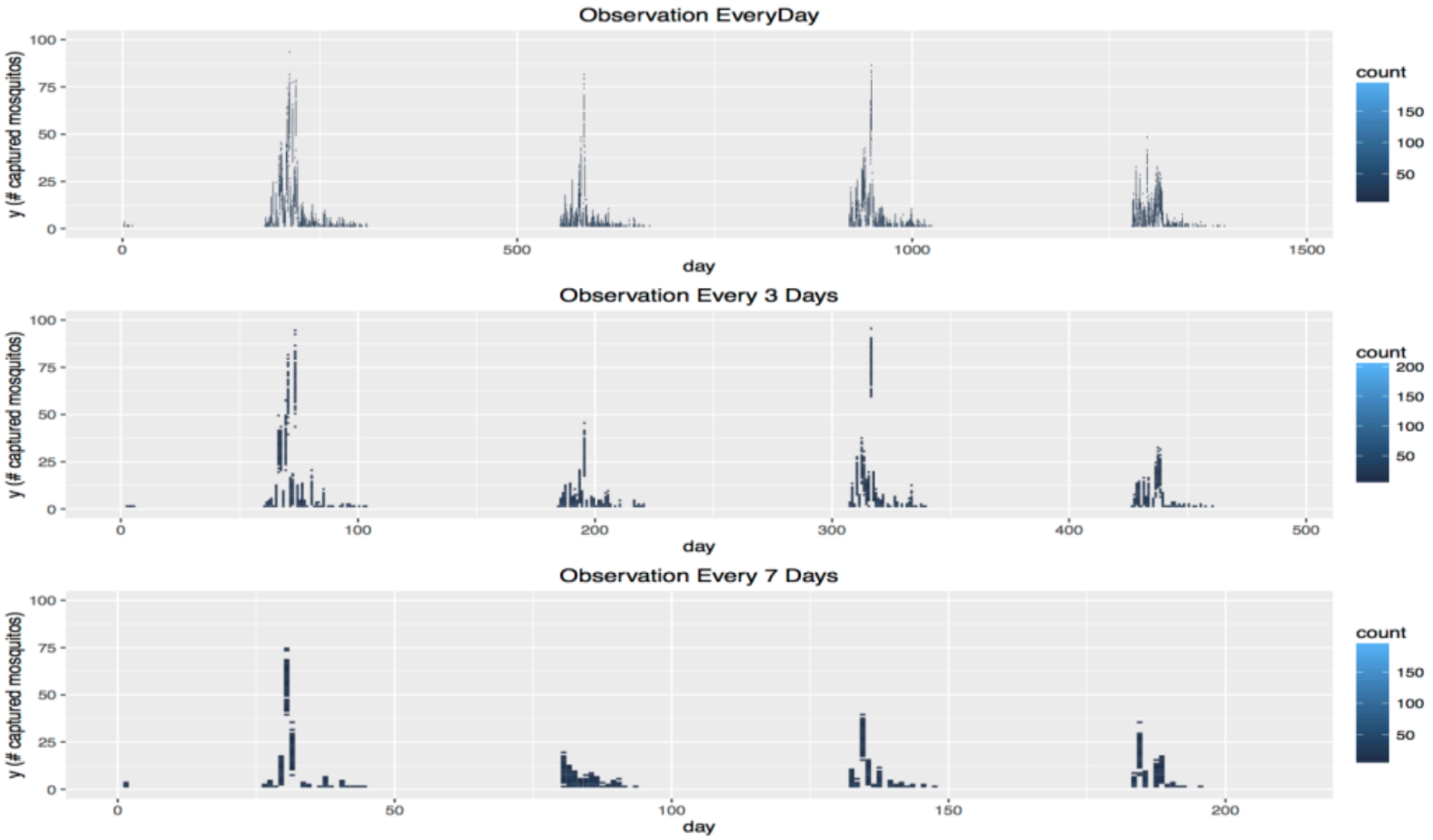

Figure 4.10: Mosquito counts in three scenarios

When observing scenario 1, we can see that the points on the graph are very tightly knit. This indicates a continuous stream of information. However, in scenario 3, we can see that there are more frequent breaks and longer distances separating the points; this can be caused by the following reasons: (1) an immediate, abrupt change in mosquito population or other environment variables; (2) the sampling period associated with the mosquito time series increases from 1 day to 3 days and 7 days; this indicates that there is some loss of information in the mosquito population. Due to the fact that we observe related variables which affect the size of the mosquito population once a day under scenario 1, the quantity of information is large and of high frequency. However, under scenario 3, we observe these variables once per seven days; the quantity of information is small compared with scenario 1. The loss of information may or may not be very important for this study.

We also can observe the four peaks in scenario 1 are higher than the four peaks in scenario 3. This is because the related data pertaining to the population size is essentially averaged over 7 days rather than varying on a per day basis.

2. Results from performing MCMC under the three different scenarios

Intuitive image contrast: The below three graphs, one for each scenario where observations for captured mosquitoes are collected every (1, 3, 7) days. For each scenario, its results are obtained from using the MCMC sampling method with two chains to sample for each of the 109 repeatedly generated datasets (indexed from 0 to 108), results in 218 (109x2) chains in total. These chains are indexed from 0 to 217 such that a pair of chains with indices $(\mathrm{i}, \mathrm{i}+1), \mathrm{i} = 0,1, \ldots, 108$ correspond to the output of MCMC sampling for generated dataset with index (i). Then for each estimated parameter in each chain, a bar was drawn to depict its HPD interval under three different scenarios. The horizontal axis denotes the parameter value (β) and the vertical axis denotes the index of a model run. The vertical red line denotes the true values for the parameters (β).

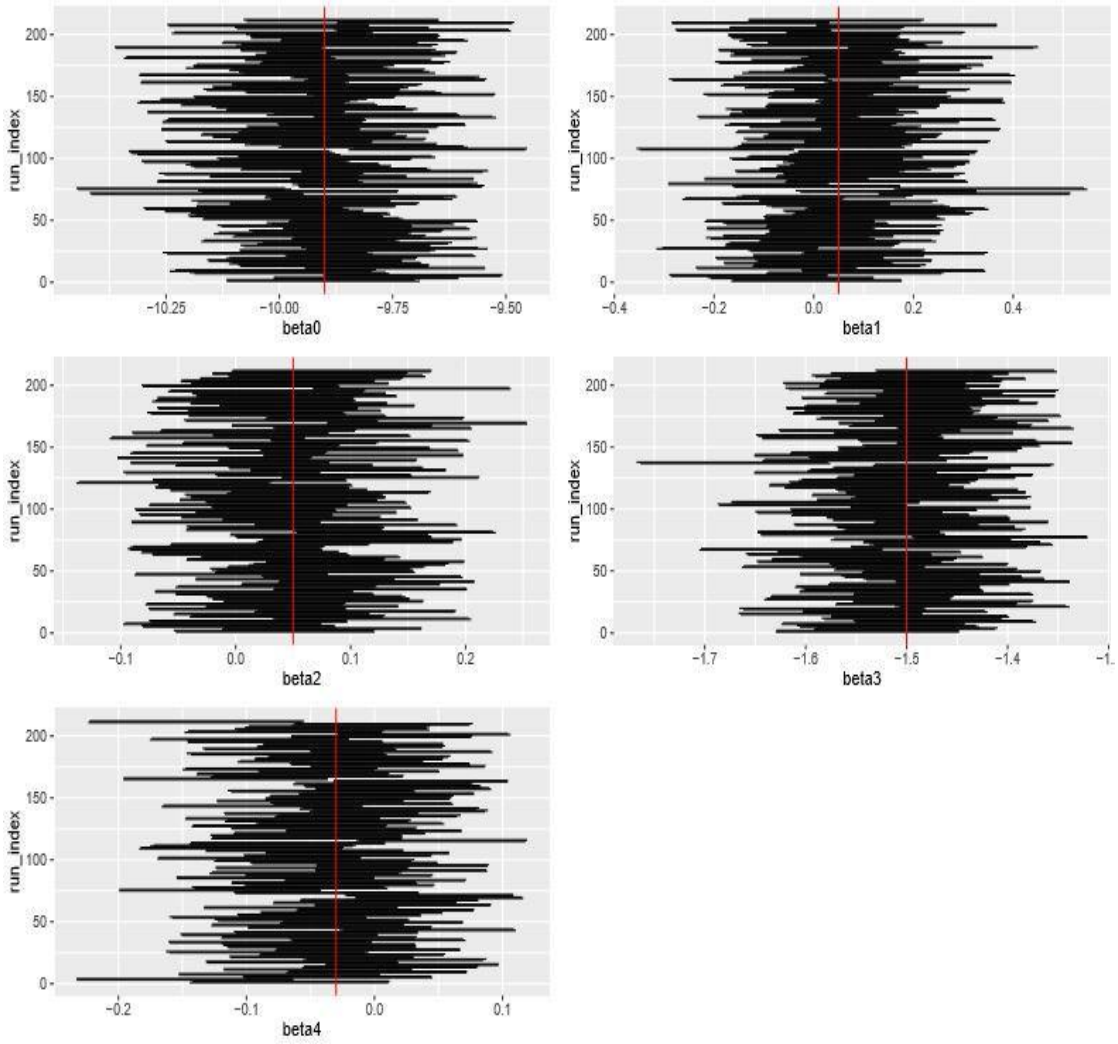

Figure 4.11: Scenario 1 sampling period based on 1 day

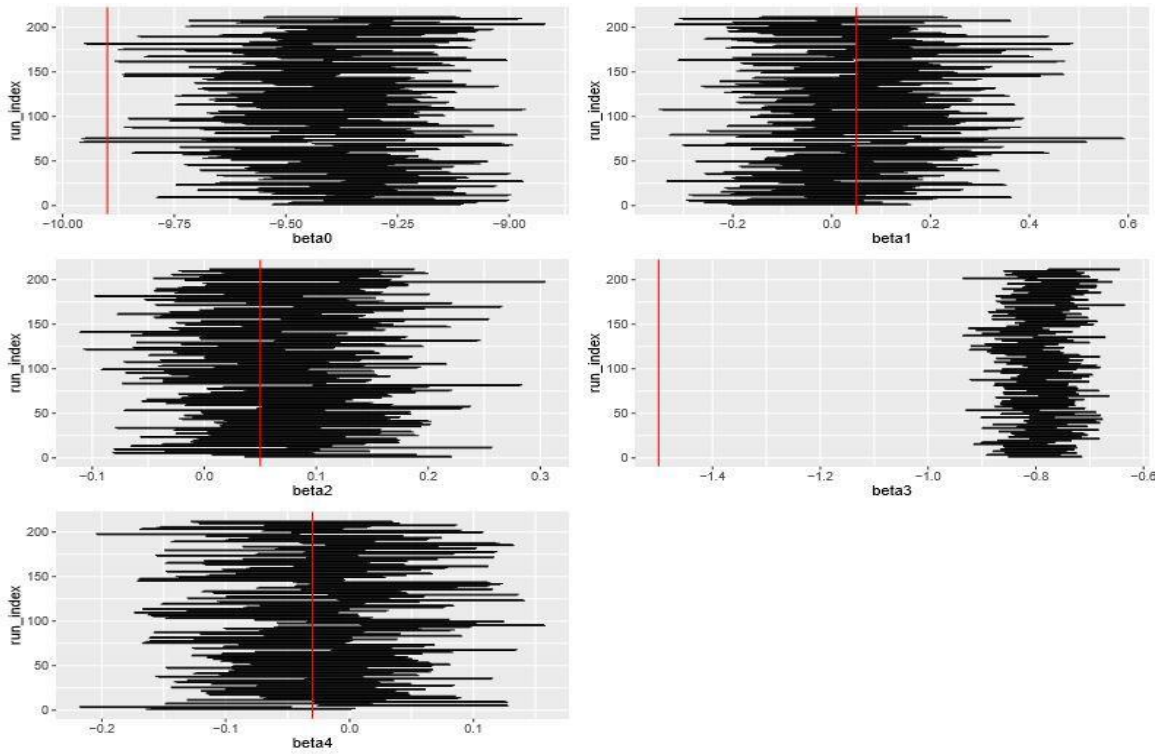

Figure 4.12: Scenario 2 sampling period based on 3days

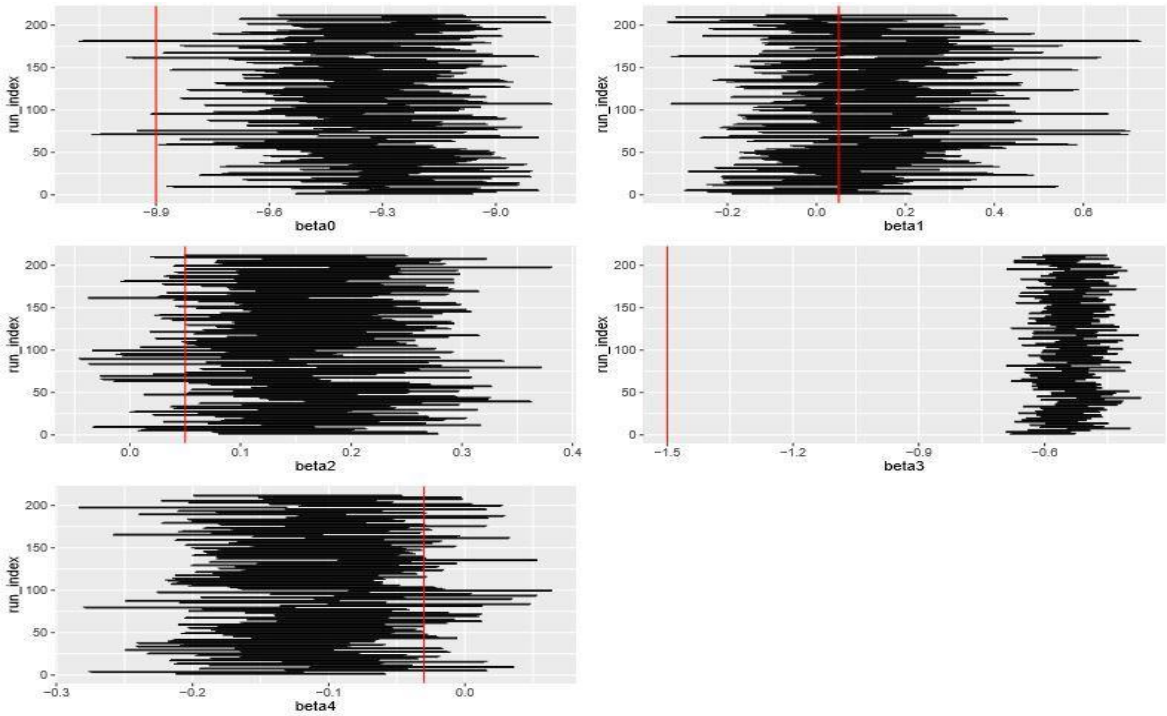

Figure 4.13: Scenario 3 sampling period based on 7 days

Comparing the above graphs, based on the same number of seeds under three different scenarios, we notice that the highest posterior density (HPD) intervals for all parameters except β_3 became wider as the sampling period for the mosquito population increased from 1 day to 3 days to 7 days.

Overall, we can see that all of the true values are almost completely covered by the HPD intervals under scenario 1; however, the true values for β_0 and β_3 are almost completely uncovered by the HPD intervals under scenarios 2 and 3.

Quantitative analysis: The Table 4.1 is a simple statistical analysis of the number of runs from the coverage rate, which is the count of times that the HPD intervals ($\alpha = 5\%$) cover the true value of parameters β divided by total count of 109 runs, under three scenarios. The results of these analyzes are consistent to those of the visual images.

Table 4.1: Coverage rate of the HPD intervals under three scenarios

Parameters true value	**Ratio for 1-day**	**Ratio for 3-day**	**Ratio for 7-day**	**Explanation for β**
$\boldsymbol{\beta_0 = -9.9}$	93.58%	2.75%	7.34%	Intercept
$\boldsymbol{\beta_1 = 0.05}$	90.83%	88.99%	89.91%	Coefficient of Tem.
$\boldsymbol{\beta_2 = 0.05}$	96.33%	92.66%	42.20%	Coefficient of R.H
$\boldsymbol{\beta_3 = -1.5}$	95.41%	0.00%	0.00%	Coefficient of W.S
$\boldsymbol{\beta_4 = -0.03}$	91.74%	90.83%	50.46%	Coefficient of Pre.

The explanations of the results: (posterior distribution: $\pi(\beta|D)$)

(1) Under 1-day sampling, the coverage rates for all parameters exhibit a satisfactory performance of over 90.00%. This may indicate that the information for the environment

variables which is more realistically reflecting actual state is continuous and smooth. In other words, the distortion of the information is relatively small.

(2) Under the 3-day and 7-day sampling regimes, we have noticed that the coverage rates decreased for β_2, β_3 and β_4 as the sampling period for the mosquito population increased from 1 day to 3 days to 7 days. We have also noticed that the coverage rates for β_0 and β_3 are very low, with the coverage rate for β_3 notably being zero. This phenomenon may be caused by the following reasons:

a) We use the average of the information of environmental variables. This average of information may lead to serious distortion relative to the actual information of a particular day. For example, suppose that under 3-day scenario the windspeed of first day and second day are 1mph and 3mph, respectively. At that windspeed, mosquitoes are more active and conductive to their reproduction. This means that there is more chance to capture mosquitoes under this condition. But suppose further that on the third day the mean windspeed reaches 20mph, and that mosquitoes cannot function at this windspeed. Right now the average of the windspeed across the three days is 8mph. At this windspeed, the activities of mosquitoes completely differ from the first day and second day condition because mosquitoes cannot tolerate windspeed higher than 7mph [37]. This means that the really useful information about first day and second day is highly distorted by using the average value. This distortion will cause an increased error.

b).The probability of capturing a mosquito is very sensitive to change of windspeed, as is known from common sense. A fluctuation can in some cases lead to a big change in the probability of capturing a mosquito.

(3) At the same time, we have also noticed that the coverage rates of β_2 and β_4 are decreasing as the sampling period for the mosquito population is increasing from 1 day to 3 days to 7 days.

As discussed above with respect to average of windspeed, use of the average of precipitation (rainfall) over multiple days can also lead to a distortion of the actual information about real rainfall with this distortion being in either a positive or negative direction. This may create significant effects on the estimated probability of capturing a mosquito. Especially for 7 days, this ratio appears to be significantly reduced from 91.74% daily to 50.46% for 7 days.

(4) Another key issue, which may affect the above results, relates to is nonlinearity in our logistic regression model. There are two functions involved in this; one is the logistic function, another is the link function, characterized below in Eq4.4 and Eq4.5, respectively:

$$p = \frac{e^{\eta(x)}}{1+e^{\eta(x)}} \quad \text{.....(Eq 4.4)}$$

and

$$\eta(x) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \beta_4 * x_4 \quad \text{.....(Eq 4.5)}$$

where $x_i(i = 1,2,3,4)$ denote the environmental variables. Under scenario 2 and scenario 3, we apply the average of information of environmental variables in the right side of link function (Eq 4.5). Due to the nonlinearity of the logistic function, after the transformation of $\eta(x)$ given in (Eq 4.4), the results differ from the corresponding average of the probabilities resulting from considering each day independently. This reflects the more general fact that in the case of a non-linear function, function applied to the average of a distribution is not the same as the average of the function applied to samples from the distribution. In other words, that is, the fitted model is not the same as the data generating model for 3-day and 7-day data.

Overall, the distortion of information is hypothesized to cause the above phenomena. Finally, the sensitivity analysis suggests that the result of scenario 1 is better than results of scenario 2 and scenario 3.

Chapter 5 Discussion and Conclusion

5.1 Discussion

The main purposes of this study are to investigate how the proposed computational method (combining System Dynamics modeling and MCMC) works and to evaluate the performance of this method when it is applied at three different sampling frequencies of observed data (1-day, 3-day and 7-day). Based on these goals, we designed the following simulation experiments: we applied System Dynamics modeling and a Logistic Regression model to generate synthetic observations of the number of captured mosquitoes under a daily scenario, and applied Bayesian inference to analyze the simulated data. After investigating the results of the generative model based on daily weather information, we have found that the amount of the synthetic captured mosquitoes (y_i) changed with the change in weather variables (empirical data), which offers confidence that the results are reflective of the known actual situation assumed in the simulation experiment. Figure 5.1 shows the change of number of generated mosquitoes (y_i) in a time series running from 2010 to 2013.

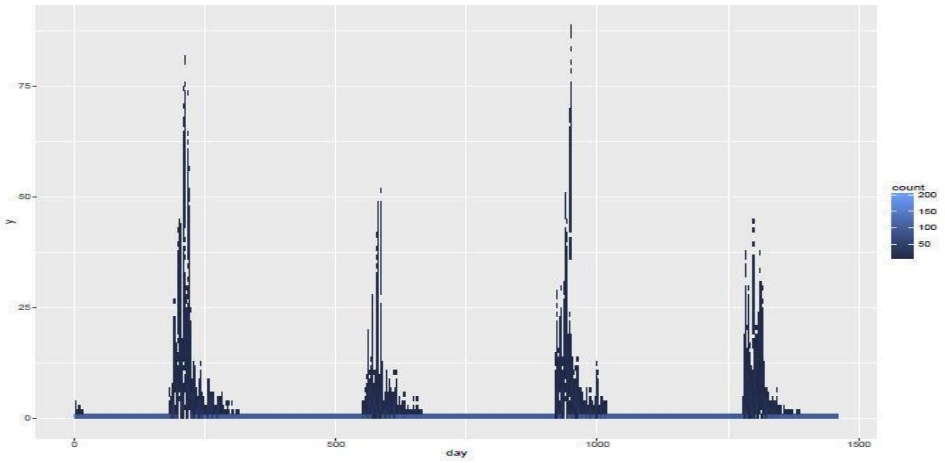

Figure 5.1 The change of number of generated mosquitoes (y_i) from 2010 to 2013

The number of mosquitoes fluctuates with changes in environmental variables; especially for the summer months, the amount of mosquito population has an obvious increase which causes a peak in captured mosquitoes each summer.

In order to make an inference(s) about the parameters (β), consisting of the intercept and the coefficients of the link function in the logistic regression model (see chapter 4 section 4.2.2) and exploring how changes in information about the environmental variable impact on the parameters (β), we constructed a Bayesian Analysis model by using the Rjags framework for MCMC. From a statistical point of view, the results of the MCMC suggested that the distribution of these parameters (β), which is a posterior distribution, follows a normal distribution based on the density plots. Also note that Markov chains sampling from the posterior distribution for the parameters (β) converge to a target distribution based on the trace plots. These results suggest that the performance of this proposed computational method is satisfactory.

We also carried out a sensitivity analysis in Chapter 4 based on changing the sampling period for the mosquito population (1-day, 3-day and 7-day). We applied the Highest Posterior Density (HPD) interval method to summarize some important statistical features of the posterior distribution for parameters $\beta_i (i = 0,1,2,3,4)$ under three different frequencies. The results of HPD interval diagnosis have shown that the HPD interval (at level_$\alpha = 5\%$) covers the true values for all β over 90% (see Chapter 4, table 4.1) based on the everyday sampling scenario. Considering the results of experiments under 3-day and 7-day sampling compared with the results of 1-day sampling in table 4.1, it appears to suggest that higher frequency of sampling the mosquito population yields a higher accuracy of our model. Also the fraction of the true values β covered by the HPD interval decreased as the frequency of sampling drops. This result seems to be somewhat unsatisfactory from its appearance alone. However, in reality, the result truly reflects the following facts. We know that the data generating system is based on daily information and thus the "plus-in" average values of independent variables (environmental variables) are not the true covariates in the model. In fact, if we plug in the average values of environment variables on the right hand side of logistic regression model, the left hand side should be the average values of logit (probability of capturing mosquitoes). But when we fit the logistic model to average values of environmental variables, we are implicitly assuming the left hand side to be logit (average values of the probabilities over three days or seven days). Therefore, the fitted model is not the same as the true data generating system.

Based on the rationality and feasibility (operability) of the proposed computational method (combining System Dynamics modeling and MCMC), the output results, which depend on different frequencies (1-day, 3-day and 7-day) into the models, are compared and analyzed. In the process, we note that as the input frequency increases, the accuracy of the model output results is improved. This means that the accuracy of output results of the model rises with the frequency of the input data.

The results with inputting of the high frequency data are better than the results based on the low frequency data input. This raises a practical challenge in the real world. Generating a high frequency sample is simple in simulation experiments. But in the real world, the generation of a high frequency data sample involves many aspects, the most important of which is the cost.

In the real world, improving the accuracy of the model means that the measurement cost will be increasing. Therefore, a higher accuracy of the model does not necessarily mean a better scenario. An ideal accuracy of the model should meet the following two principles: (1) “sufficient principle”, which is the accuracy of the model satisfies the requirements of practical operation; (2) a minimum cost principle. The key here is to find a balance point, which is an ideal (or practical) frequency, to balance the two principles mentioned above. Determination of the optimal sampling frequency of the observed data in practice lies beyond the scope of this study.

5.2 Potential Future Work

5.2.1 Important Variables

Our result section has showed that the generated model assumed that environmental variables have a significant impact on both of the size of mosquito population and probability of capturing a mosquito (see chapter 4 section 4.1.3). Due to the purpose of this study, this thesis does not primarily focus on exploring both of such sides of the impact of the following environmental variables.

Temperature: In this study, temperature is noted to have a significant impact on both the size of the mosquito population as well as on the probability of capturing mosquitoes. Studies

have shown that in reality, within a certain temperature range, there is a positive relation between amounts of mosquito population and temperature. However, when temperature rises beyond this range, there is a negative relation [38]. Both the System Dynamics model and Probability Generating model are conditioned on the temperature range where the amount of mosquito population and temperature are positively correlated in our study.

We need to emphasize that the temperature which is selected by this study refers to the air temperature. Most of the researchers used air temperature or soil temperature in their research to explore the influence of temperature on dynamic change of the mosquito population. But water temperatures seem to be associated more closely with early aspects of the mosquito's ecosystem cycle (e.g., larval and pupal stages) in our research area. In particular the water temperatures in surrounding area of the capture mosquito sites could help complement the information available for air temperatures. Due to the existence of a non-linear relationship (generally a big change in air temperature will lead to a small change in water temperature) between air temperature and water temperature, considering water temperature rather than air temperature would lead to the immature development stage of some mosquitoes species being shortened by 4-11 days [39]. Analyzing and comparing the differences on mosquito population under the two temperatures regimes could be an important priority for future study.

Relative Humidity: Relative humidity plays a crucial role in affecting the mosquito population [39]. As common sense would suggest, low of levels of relative humidity will shorten the lifespan of mosquitoes. However, this study did not explore the potential indication of relative humidity on availability of breeding pools, which may reveal indirective impacts of relative humidity on mosquito egg and larvae develop to adult mosquito population. It could also be considered for further study.

Precipitation: Precipitation is a key fact in creating and maintaining suitable larval habitats; excess precipitation will destroy the development environment of larva and significantly affects the size of mosquito population, as well as affecting the probability to capture a mosquito. Therefore, there is also an influence relative to both sides (increasing or decreasing) of the probability of capturing a mosquito.

Wind speed: Wind speed has an important impact on the probability of capturing a mosquito. In this study, we assumed an identical wind speed for the study area. However, a wind speed near capture sites seems more important than regional wind speed. So we suggest using the wind speed near capture sites in future studies.

Missing important environmental variables: In addition to the above environmental variables that we have mentioned and which were included in our model, our model also neglects some important environmental factors which have a significant influence on the size of the mosquito population. For example, the following environmental factors should be considered for inclusion:

(1) The number of sunlight hours available has been identified as playing an important role in each stage of mosquito development [23].

(2) The evaporation rate, which directly impacts the moisture content of the soil surface which is a key factor for egg, larva and pupa stages of mosquito development [40].

Finally, our study leaves above suggestions and environmental factors for future consideration.

5.2.2 Interactions of Independent Variables

Temperature, Relative humidity, Windspeed and Precipitation were selected as environmental factors which have a significant effect on mosquito population and probability of capturing a mosquito. In our model, we only consider these factors' main effects on the dynamic change of mosquito population and probability of capturing a mosquito, and neglect the complex interactions among these environmental variables. For instance, high temperatures and windspeed often lead to decreases the relative humidity and precipitation; interaction factors reflecting such dependencies could serve to offset the positive effects of temperature on the size of mosquito population. The effects of these interactions on the size of mosquito population and probability of capturing a mosquito have yet to be incorporated into our model, but we should aware that these could be important to fully understanding the reasons of dynamic changes of mosquito population. This task is reserved for future research work.

5.3 Conclusion

In this thesis, we have investigated a good estimate of mosquito population with the combination of real-world data (environmental variables) and synthetic data. We have found that insights from real-world data (environmental variables) can be secured via MCMC of a simpler, well-structured probability generating model when securing findings from a combination of a reasonably complex system dynamics model and probability generating model. Additionally we have revealed the importance of sampling the mosquito population every day for reliably estimating parameter values, rather than pursuing the standard approach of sampling the mosquito population every week. Such work offers to inform prediction and control of mosquito-borne diseases in future transmission.

In the future, we should focus on the following : (1) including some new environmental variables in our model which have significant impact on the abundance of mosquito population and probability of capturing a mosquito; (2) instead of air temperature and regional windspeed, consider water temperature and windspeed nearby capture (trap) sites; (3) considering the interaction effect of environmental variables on the mosquito population size and probability of capturing a mosquito in the model; (4) extending the time span of data about the environmental variables from a current 4 years to 10 years.

While these ideas have received little attention to date, the process of assessing and quantifying not only enables us to better verify the explanatory power of our models, but also to produce more reliable forecasts of future spatiotemporal patterns of WNv transmission.

Both statistical and computer models have important roles to play in simulation synthesis of observed mosquito ($counts\ y_i$). In Bayesian inference, we used the MCMC method as implemented via the Rjags tool to draw samples for parameter values from the posterior distribution: this approach exempts us from formulating and solving problems in closed form. As a result, it broadens the scope of problem-solving in the study field. It is hoped that the MCMC method provides a basis for future modelling efforts within the field.

In spite of the fact that it has long been understood that a range of environmental variables, including temperature, relative humidity, windspeed and precipitation, etc., have a significant impact on abundance of mosquito population and on probability of capturing a mosquito, there has been little work offering an analysis to this effect. By developing some mechanistic models which incorporate the effects of average temperature, average relative humidity, average windspeed and average precipitation on abundance of mosquito population and on the probability of capturing a mosquito, we have taken important assumptions in which average temperature and average relative humidity are associated with a positive relation with the probability of capturing a mosquito, and average windspeed and average precipitation exhibit a negative relation with the probability of capturing a mosquito. We tested these assumptions on synthetic data by using MCMC (as implemented in Rjags) and HPD diagnostic with an accuracy rate over 90% (at α=5%) under daily situation; and quantitatively analyzed these relations, which appear to be well-characterized by normal distributions. This study helps to provide a framework for possible future analysis of empirical data, and raises concerns about the inadequacy of existing data to sufficiently resolve the relationships involving environmental variables.

In view of the fact that the complexity of the factors affects the mosquito population and probability of capturing a mosquito in the real world, as a result, our model is both simplified and lopsided compared to a real world scenario. The simplicity and one-sidedness are due to the fact that our data sets only cover four years and only consider a limited number of environmental factors (temperature, relative humidity, windspeed and precipitation) in our model. Therefore, it is hard to say how accurately our model truly reflects a real world scenario. It is also difficult to assess the significance of the results and conclusions from running this model; but in any case, the meaning and purpose of this study is to provide foundational ideas and a framework for future endeavors in this field.

References

[1]. Brown, L. (1993). *The New shorter Oxford English dictionary on historical principles*. Oxford [Eng.]: Clarendon. ISBN 0-19-861271-0.

[2]. Smithburn, K., Hughes, T., & Burke, A. (1940). A neurotropic virus isolated from the blood of a native of Uganda. *American Journal of Tropical Medicine1940, 20, 471-492.* Retrieved from https://www.cabdirect.org/cabdirect/abstract/19412700112?freeview=true

[3]. Hayes, C. G. (2001), West Nile virus: Uganda, 1937, to New York City, 1999. *Annals of the New York Academy of Sciences*, 951: 25–37. doi:10.1111/j.1749-6632.2001.tb02682.x

[4]. Nash, D., Mostashari, F., & Fine, A. (2001). The outbreak of West Nile virus infection in the New York City area in 1999. *The New England Journal of Medicine,* 344:1807-1814. doi: 10.1056/NEJM200106143442401

[5]. West Nile virus: causes, symptoms, treatment, diagnosis. (http://chealth.canoe.com/condition/getcondition/West-Nile-virus.)

[6]. Government of Canada. 2015. Surveillance of West Nile virus: Human Surveillance. (http://healthycanadians.gc.ca/diseases-conditions-maladies-affections/disease-maladie/west-nile-nil-occidental/surveillance-eng.php)

[7]. Government of Saskatchewan. West Nile virus (WNv) Surveillance Results and Transmission Risk 2015 (https://www.**saskatchewan**.ca/.../**westnilevirus**serveillancetransmissionsept%2010%202...)

[8]. Corrigan, R. L., Waldner, C., Epp, T., Wright, J., Whitehead, S. M., Bangura, H., . . . Townsend, H. G. (2006). Prediction of human cases of West Nile virus by equine cases, Saskatchewan, Canada, 2003. *Preventive Veterinary Medicine, 76*(3-4), 263-272. doi:10.1016/j.prevetmed.2006.05.008

[9]. Schellenberg T.L., Curry P.S., Anderson M.F., Campbell C.A., et al (2006). Seroprevalence of West Nile virus in Saskatchewan's Five Hills health region, 2003. Can J Pub Health 2006; 97:369–373.

[10]. Téllez-Zenteno, J. F., Hunter, G., Hernández-Ronquillo, L., & Haghir, E. (2013). Neuroinvasive West Nile Virus Disease in Canada. The Saskatchewan Experience. *The Canadian Journal of Neurological Sciences, 40*(04), 580-584. doi:10.1017/s0317167100014700

[11]. West Nile virus transmission cycle. (http://www.mayoclinic.org/West-Nile-virus-transmission-cycle/img-20006044)

[12]. American Mosquito Control Association: Mosquito-Borne diseases.(2014) (http://www.mosquito.org/mosquito-borne-diseases)

[13]. USA. Illinois Department of Public Health: Mosquitoes and Disease. (2007,March 29) (http://www.idph.state.il.us/envhealth/pcmosquitoes.htm)

[14]. Reiter, M.E. LaPointe, D.A. (2007). Landscape Factors Influencing the Spatial Distribution and Abundance of Mosquito Vector Culex quinquefasciatus (Diptera: Culicidae) in a Mixed Residential–Agricultural Community in Hawai'i. J*ournal of Medical*

Entomology,44(5), 861-8.

[15]. U.S. Department of Health and Human Services Public, Public Health Service, Centers for Disease Control and Prevention National Center for Emerging and Zoonotic Infectious Diseases, Division of Vector-Borne Diseases: West Nile Virus in the United States: Guidelines for Surveillance, Prevention, and Control. (2013, June 14.) Retrieved from http://www.cdc.gov/westnile/resources/pdfs/WNvGuidelines.pdf

[16]. News, C. (2015, June 26). West Nile virus season beginning in Saskatchewan. Retrieved August 31, 2017, from http://www.cbc.ca/news/canada/saskatchewan/west-nile-virus-season-beginning-in-saskatchewan-1.3128940

[17]. Barker, C. M., Eldridge, B. F. & Reisen, W. K. (2010). Seasonal Abundance of *Culex tarsalis* and *Culex pipiens* Complex Mosquitoes (Diptera: Culicidae) in California. *Journal of Medical Entomology, 47*(5), 759–768.

[18]. Cianci, D., Van den Broek, J., Caputo, B., Marini, F., Della Torre, A., Heesterbeek, H., et al. (2013). Estimating mosquito population size from mark-release-recapture data. *Journal of Medical Entomology, 50(3),* 533–542. doi: 10.1603/ME12126.

[19]. Dowdeswell, W. H., Fisher, R. A., and Ford, E. B. (1940). The quantitative study of populations in the lepidoptera I. Polyommatus icarus rott. *Annals of Eugenics,* 10(1), 123–136. doi: 10.1111/j.1469-1809.1940.tb02242.x

[20]. Villela, D. A. M., Codeço, C. T., Figueiredo, F., Garcia, G. A., Maciel-de-Freitas, R., &

Struchiner, C. J. (2015). A Bayesian Hierarchical Model for Estimation of Abundance and Spatial Density of Aedes aegypti. *PloS One*, 10(4), e0123794. http://doi.org/10.1371/journal.pone.0123794

[21]. Shaman, J., Stieglitz, M., Stark, C., Le Blancq, S., Cane, M. (2002). Using a dynamic hydrology model to predict mosquito abundances in flood and swamp water. *Emerging Infectious Diseases, 8(1),* 6–13. doi: 10.3201/eid0801.010049.

[22]. Brailsford, S. C. B., Roberto; Angelis, Vanda De; Mecoli, Mariagrazia. (2009). System Dynamics Models to Assess the Risk of Mosquito-Borne Diseases and to Evaluate Control Policies. *[Congresses (Conference Proceedings)]. Proceedings of the 35th International Conference on Operational Research Applied to Health Services (ORAHS), 1979(January 2007),* 1-10.

[23]. Theoret, C., Richards, M. (2014). Mosquito Population Model. *University of Saskatchewan computer science. Course CMPT 394 term project, instructed by* Professor Nathaniel Osgood.

[24]. (n.d.). Retrieved June 12, 2016, from http://www.pmcorp.com/Products/Simulation/AnyLogic.aspx?gclid=Cj0KEQjw9tW5BRDk29KDnqWu4fMBEiQAKj7spwaqkMR7d548ZgDMMRja-m7HzNRKvUITPy15TqyhR5UaAuLs8P8HAQ.

[25]. (n.d.). Retrieved June 12, 2017, from http://www.anylogic.com/anylogic/help/

[26]. Pruyt, E. (2013). *Small system dynamics models for big issues: Triple jump towards real – world dynamic complexity*. Delft City, The Netherlands: TU Delft Library, Delft, The

Netherlands.

[27]. Albin, S., Forrester, J.(1997). *Building a System Dynamics Model*. Prepared for the MIT System Dynamics in Education Project. Massachusetts Institute of Technology. Cambridge, Massachusetts, United States.

[28]. Brooks, S., Gelman, A., & Jones, G. L. (2011). Handbook of Markov chain Monte Carlo. Boca Raton: CRC Press.

[29]. Gamerman, D., Lope, H. F. (2006). *Markov Chain Monte Carlo Stochastic Simulation for Bayesian Inference*. Second Edition. 6000 Broken Sound Pkwy NW, Boca Raton, FL 33487, USA: Chapman & Hall/CRC Taylor & Francis Group.

[30]. Andrien, C., Freitas, N. D., Doucet, A., Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning, 50(1-2),* 5-43. doi (10.1023/A:1020281327116).

[31]. Givens, G. H., & Hoeting, J. A. (2013). *Computational statistics*. Oxford: Wiley-Blackwell

[32]. Plummer, M., Stukalov, A., Denwood, M. (2016, February,20). Bayesian Graphical Model using MCMC (Package 'rjags'). Retrieved from https://cran.r-project.org/web/packages/rjags/rjags.pdf.

[33]. Weicheld, J. J. (2015). Impact of Environmental Factors on Mosquito Population Abundance and Distribution in King County, Washington. *CEUR Workshop Proceedings 1542(9),* 33-36. doi (10.1017/CBO9781107415324.004).

[34]. Environment Canada, (2015, May 25). Saskatchewan - Weather Conditions and Forecast by Locations, Retrieved February 12,2016, from https://weather.gc.ca/forecast/canada/index_e.html?id=SK

[35]. Chen, M. H., Shao, Q. M. (1999). Monte Carlo Estimation of Bayesian Credible and HPD Intervals. *Journal of Computational and Graphical Statistical*, 8(1), 69-92. DO (10.2307/1390921)

[36]. SAS/STAT(R)9.2 User's Guide,Second Edition. (2010, April 30). Retrieved November 19, 2016, from https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_introbayes_sect005.htm#statug.introbayes.bayesinterval

[37]. What You Need To Know About Mosquitos To Not Have Them. (n.d.). Retrieved September 02, 2017, from https://www.spalding-labs.com/products/mosquito_control_products/mosquito_torpedo/p/what_you_need_to_know.aspx

[38]. Abiodun, G.J., Maharaj, R., Witbooi, P. et al. (15 July 2016). Modelling the influence of temperature and rainfall on the population dynamics of *Anopheles arabiensis. Malaria Journal,15*(1), 364. DOI**:** 10.1186/s12936-016-1411-6

[39]. Waldock, J., Chandra, N. L., Lelieveld, J., Proestos, Y., Michael, E., Christophides, G., & Parham, P. E. (2013). The role of environmental variables on *Aedes albopictus* biology and chikungunya epidemiology. *Pathogens and Global Health*, *107*(5), 224–241. http://doi.org/10.1179/2047773213Y.0000000100

[40] Gong, H., Degaetano, A., Harrington, L.(2007). A Climate Based Mosquito Population Model. *Proceedings of the World Congress on Engineering and Computer Science 2007, WCECS 2007, October 24-26,2007, San Francisco, USA.*